GENOMICS OF POPULATION DECLINE IN THE FLORIDA SCRUB-JAY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Nancy Chen

August 2014

GENOMICS OF POPULATION DECLINE IN THE FLORIDA SCRUB-JAY

Nancy Chen, Ph.D.

Cornell University 2014

Myriad species are experiencing declining numbers worldwide, yet many details of the eco-evolutionary responses to population decline remain poorly characterized in wild species. Rapidly shrinking populations experience loss of genetic diversity and increased homozygosity from changes in the balance of genetic drift, mutation, gene flow, selection, and inbreeding. Despite good theoretical knowledge of the impact of these factors on population genetics, thorough empirical evaluations in natural populations are scarce because they demand huge field and laboratory investments. Recent development of next-generation sequencing technologies now permits discovery and genotyping of large numbers of genetic markers in any species. Combining genomic data with long-term demographic and pedigree data from a natural population allows us to explore fundamental questions concerning the population genetic consequences of declining population size. Here, I describe two bioinformatics methods for analyzing genomic data and apply these methods to develop substantial genomic resources for one of the longest-studied endangered species in the world, the Florida Scrub-Jay (*Aphelocoma coerulescens*). One method identifies sequences specific to the heterogametic sex chromosome, and another uses pedigree information to improve single nucleotide polymorphism (SNP) discovery. A population of individually banded Florida Scrub-Jays at Archbold Biological Station has been studied for more than 43 years, providing an unparalleled model for research on the genomics of population decline. I geno-

typed 3,578 individuals sampled through time at 15,416 genome-wide SNPs. To investigate the impact of regional population decline on our stable study population, I used 7,404 autosomal SNPs to calculate levels of heterozygosity and inbreeding. Decreasing immigration over time was correlated with an increasing mean inbreeding coefficient of the birth cohort, and inbreeding was correlated with higher rates of hatching failure. These results imply that despite the small and shrinking size of peripheral populations, their small but measurable genetic differentiation from the central population may give them a crucial role in maintaining genetic diversity. This study, which marks the beginning of a detailed longitudinal investigation of genomics in a wild animal population, underscores the vital importance of maintaining gene flow among remnant populations, for the Florida Scrub-Jay specifically, and for other declining species more broadly.

BIOGRAPHICAL SKETCH

Nancy Chen was raised in Yorba Linda, CA. She had a love of animals and nature from an early age, and her parents routinely took her to the Los Angeles Zoo, the California Science Center, and many National Parks. Nancy's interest in science was encouraged by an excellent seventh grade science teacher, Mrs. Lane, who suggested she explore different fields of science by joining the Science Olympiad Team. Science Olympiad ended up playing a major role in her time at Troy High School, and it is through this competition that Nancy first developed an interest in ornithology. Nancy got her first research experience the summer before senior year of high school, when she studied the structure of the DNA backbone in Professor Fu-Ming Tao's lab at California State University, Fullerton. After graduating in 2003, Nancy moved to Harvard University, where she majored in Biochemical Sciences because she couldn't decide if she liked biology or chemistry more. Her first two summers were spent doing chemistry research in California: she analyzed the thermoelectric properties of a semiconductor under Dr. G. Jeffrey Snyder at the Jet Propulsion Laboratory in 2004 and used NMR to study conformations of *meso*-tartaric acid under Professor John D. Roberts at Caltech in 2005. Spring semester sophomore year, Nancy took an ornithology class taught by Professor Scott V. Edwards, who introduced her to evolutionary biology and changed her life. Nancy joined the Edwards lab the beginning of her junior year, and there she completed an undergraduate honor's thesis on population genetics of House Finches. Nancy graduated from Harvard *magna cum laude* with an A.B. in Biochemical Sciences in 2007. In Fall 2007, Nancy enrolled in the PhD program at Cornell University, where she is pursuing her interests in avian evolutionary genomics under the guidance of Professor Andrew G. Clark and Professor John W. Fitzpatrick.

To my family and all the wonderful mentors in my life.

ACKNOWLEDGMENTS

This dissertation would not have been possible without the help of many people. First I'd like to thank my committee for their patience and guidance. I really benefited from the synergy of my two amazing co-advisors, John Fitzpatrick and Andy Clark. They provide completely different perspectives on my work, but they are both eternal optimists and big dreamers. Together we started this ambitious project, which has been so much fun and will likely form the basis of my entire career. Thank you, Fitz and Andy, for trying something new. Rick Harrison has essentially been a third advisor to me, and I believe that adding him to my committee was one of the best decisions I made during my graduate career. Thank you for your honest and invaluable advice, and your unwavering faith in my abilities. My final committee member is Ton Schat, who taught me everything I know about avian immunology and has helped me think through minute details of all my field experiments (which unfortunately mostly failed, but I hope to revisit those ideas soon).

I also had many mentors besides my committee. Scott Edwards was my undergraduate thesis advisor who convinced me to go to graduate school and has remained a close friend and mentor through the years. Irby Lovette welcomed me into his lab and provides a lot of crucial logistical support with the Florida Scrub-Jay samples. Andre Dhondt treated me like one of his own my first year and spent hours talking about science with me. Monica Geber coached me through some different NSF proposal writing and has generally served as an inspiration and role model.

Winston Bellott and David Page were my collaborators for chapter 2, and Cris Van Hout and Srikanth Gottipati were my collaborators for chapter 3. A large part of my research depends on a long-term demographic study, and none

of that work would be possible without the hard work and dedication of Reed Bowman, Raoul Boughton, Shane Pruett, Laura Stenzler, and many students, interns, and staff at Archbold Biological Station. I got a ton of help with labwork from many people, including Jen Grenier, Charlotte Acharya, Laura Stenzler, Amanda Manfredo, Peter Schweitzer, Grace Chi, and Rob Elshire. I'm grateful to Alex Coventry, Wes Hochachka, Cris Van Hout, Haley Hunter-Zinck, Angela Early, Tim Connallon, and Rob Unckless for statistical and data analysis advice. Thank you Qi Sun, Jarek Pillardy, Rob Bukowski, and other staff at the Cornell Statistical Consulting Unit and Cornell Center for Advanced Computing for supporting great computational resources. I was lucky to be a part of many lab groups, and I would like to thank past and current members of the Fuller Evolutionary Biology lab, the Dhondt/Dickinson lab, and the Harrison lab for many stimulating lab meetings. I especially want to thank the many members of the large and diverse Clark lab for inspiring me to become a better scientist. I have learned so much from you all. I would like to thank many administrative staff in EEB, MBG, and the Lab of O, especially Lori Beyea-Powers and Melissa Switzer, for helping me keep track of funding and paperwork and other little but necessary details.

My graduate research was funded by two NSF grants (SGER DEB 0855879 and DEB 1257628), the Cornell Lab of Ornithology Athena Fund, a Cornell Center for Vertebrate Genomics Seed Grant, and a number of small grants (Andrew W. Mellon Student Research Award, EEB Graduate Student Research Fund, Joseph Grinnell Student Research Award, AMNH Frank M. Chapman Memorial Grant, Cornell Sigma Xi Research Award). I was supported by a Cornell Presidential Life Sciences Fellowship, a NSF Graduate Research Fellowship, and a Cornell Center for Comparative and Population Genomics Fellowship.

Finally, and most importantly, I would like to thank my family and friends for all their love and support through these years. I could not have done this without them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

WILD PEDIGREES AND GENOMICS: TESTING THE CONSERVATION GENETICS PARADIGM

1.1 Introduction

In the current global extinction crisis, unprecedented numbers of species are undergoing severe population declines around the world (Butchart *et al.* 2010). Efforts to understand and mitigate this rampant loss of biodiversity have been guided by a set of population genetic models that predict short-term eco-evolutionary responses to declining population size (the conservation genetics paradigm; Ouborg *et al.* 2006). Decreased population size is predicted to trigger a loss of genetic variation and an increase in homozygosity, which in turn will decrease both population mean fitness and adaptive potential (Frankham 2005; Willi *et al.* 2006). This prevailing model has motivated numerous basic and applied research efforts worldwide, most of which use a handful of neutral markers to assess genetic diversity and fitness in small populations (Frankham 1996; Ekblom and Galindo 2011). However, it is unclear whether variation at a few neutral markers adequately represents genome-wide variation or levels of functionally important genetic diversity (Kohn *et al.* 2006; Ouborg *et al.* 2006; Väli *et al.* 2008; Angeloni *et al.* 2012). In addition, little empirical evidence exists to support a relationship between low genetic variation, decreased population fitness, and loss of adaptive potential (Kohn *et al.* 2006). Obtaining precise measurements of neutral and functional genetic variation and understanding the mechanisms by which reduced genetic variation affects population fitness require studying the genetics of small populations at a genome-wide scale.

Stochastic processes become increasingly important with declining population size, resulting in a loss of variation and increased homozygosity (Hedrick 2001; Ouborg *et al.* 2006). Additionally, the efficacy of selection is reduced in small populations, leading to an accumulation of deleterious alleles and reduced opportunities for adaptive substitutions (Lande 1988; Lynch *et al.* 1995; Hedrick 2001). Maintaining healthy levels of heterozygosity and genetic diversity under such conditions requires (a) specific mechanisms for inbreeding avoidance, and (b) the acquisition of genetic diversity through immigration (Frankham *et al.* 2003; Kohn *et al.* 2006). Reduced genetic diversity and increased homozygosity should decrease the average fitness of the population (via inbreeding depression) and ultimately limit a population's ability to adapt to environmental change, further increasing its probability of extinction (Frankham 2005; Willi *et al.* 2006). A number of studies have used neutral markers (*e.g.*, microsatellites and amplified fragment length polymorphisms) to document decreased genetic diversity and increased inbreeding in small populations (reviews in Reed and Frankham 2003; Leimu *et al.* 2006), and their results are routinely incorporated in conservation management plans (Kohn *et al.* 2006). However, most genetic studies of declining populations assume that variation in neutral markers accurately reflects levels of functionally important genetic diversity, thereby providing a suitable proxy for fitness and the adaptive potential of populations (Ouborg *et al.* 2010; Angeloni *et al.* 2012). To date, the empirical evidence supporting this crucial relationship is limited (Kohn *et al.* 2006; Ouborg *et al.* 2006). Thus, a better understanding of the relationship between neutral and fitness-related variation is especially important for sharpening our understanding of the evolutionary processes occurring in declining populations (Kohn *et al.* 2006). More generally, a full understanding of the mechanisms by

which reduced genetic variation leads to lower population fitness requires disentangling the relative impacts of different evolutionary forces at a genome-wide scale.

Ongoing controversy regarding the utility of neutral markers to predict population fitness stems from two limitations that have inhibited definitive tests of this assumption: (1) nearly all studies use relatively small numbers of neutral markers, derived from a small fraction of the genome (Kohn *et al.* 2006), and (2) population genomic and quantitative genetic studies aimed at detecting fitness-related variation are rarely conducted in wild populations with known ecological contexts and plausible agents of selection (Stinchcombe and Hoekstra 2008). Studies conducted in a stable laboratory environment produce overly simplified models of evolutionary genetic change compared with those conducted in natural environments (Kruuk *et al.* 2000; Ellegren and Sheldon 2008). Quantitative trait loci (QTLs) and the fitness of particular genotypes can differ greatly between laboratory and field experiments (Ellegren and Sheldon 2008). Measuring individual fitness in the wild requires a substantial investment in fieldwork, with the best opportunities coming from a limited number of long term studies of marked and monitored individuals (best known pedigreed vertebrate populations are wild ungulate and songbird populations in Europe; reviewed in Kruuk 2004). Until recently, long-term studies in wild populations were severely hindered by a shortage of genomic resources (Ellegren and Sheldon 2008), but next-generation sequencing technologies now permit simultaneous examination of neutral variation throughout a genome and direct study of the genetic basis of fitness (Hudson 2008). This analysis is most informative when genomic information is combined with ecologically relevant phenotypic data (Primmer 2009).

In addition, application of genomic tools to long-term field studies can identify mechanisms that link levels of genetic diversity and heterozygosity with variation in fitness at the individual level (Ouborg *et al.* 2010; Angeloni *et al.* 2012). A key factor impacting small populations is inbreeding depression - the phenomenon of reduced fitness in inbred individuals (Frankham *et al.* 2003; Ouborg *et al.* 2006; Angeloni *et al.* 2012). Inbreeding depression is primarily caused by increased homozygosity of recessive deleterious alleles, but the role of overdominant genes is still unresolved (Charlesworth and Willis 2009). The development of next-generation sequencing tools makes it feasible to localize and perhaps describe the function of genes underlying inbreeding depression (Kristensen *et al.* 2010). Characterizing the precise genetic basis of inbreeding depression will improve our understanding of how increased homozygosity affects fitness in small populations (Ouborg *et al.* 2010).

There have been a number of reviews highlighting various issues in conservation biology that can be addressed with genomic technologies (Allendorf *et al.* 2010; Kohn *et al.* 2006; Ouborg *et al.* 2010; Primmer 2009). Here, we focus on the insights that can be gained by combining genomics with intensively studied wild populations. By overlaying genotypes onto a large pedigree and an accompanying demographic dataset, it is now possible to address fundamental questions in evolutionary and conservation genetics concerning the relative impact of drift, selection, gene flow, and mate choice on genetic variation and their consequences for population fitness. In particular, it is now feasible to empirically test the prevailing conservation genetics paradigm in the wild. A more comprehensive understanding of the theoretical underpinnings of conservation genetics will likely have broad implications for the study and management of declining populations all around the world. Below, we propose a five-part con-

servation genomics research program for comprehensively testing whether and how reduced genetic diversity predicts a decline in fitness and for providing insights into the underlying mechanism of inbreeding depression. We then provide an example of a study system that has sufficient phenotypic and genetic resources for the described approaches.

1.2 Characterize the genetic architecture of fitness-related traits

A first step in understanding the genetic response to reduced population size is the identification of genomic regions linked to fitness differences. Regions of the genome linked to fitness can be found by linkage mapping of fitness-related traits or separately testing for evidence of selection or associations with survival in candidate genes for relevant phenotypes. Once these regions are identified, their contribution to additive and dominance variance can be estimated using approaches outlined below.

1.2.1 Linkage mapping

Highly refined likelihood-based methods developed by the human genetics community can be applied to identify the genetic architecture of fitness-related traits. Linkage mapping requires both phenotypic and genotypic data for multiple individuals on a pedigree. Phenotypes can include longevity, reproductive traits such as fecundity, and morphometric measures. Marker density will depend on the mean recombination rate of the organism of interest and should be high enough to map nearly every meiotic exchange with reasonable accuracy. Heterogeneity in recombination rates across the genome means there is

varying power to detect causal loci in different regions of the genome (Nachman 2002). Therefore, the ideal marker density in a particular genomic region should take into account local recombination rates, and it may be useful to use a combination of approaches for trait mapping. Pedigree and phenotype data alone are sufficient for estimating the narrow-sense heritability (h^2) of different measured traits using mixed linear models known as the Henderson "animal model" (Henderson 1984).

Knowledge of the contribution of genetic factors in determining trait variance is crucial for understanding the evolutionary processes that shape the genetic variation underlying phenotypic traits. Statistical inference methods can determine both the location and effect size of genomic regions that harbor variant alleles affecting measured morphological or life-history traits. A commonly-used variance components method for linkage mapping is described in George *et al.* (2000). First, identity-by-descent (IBD) is estimated between all individuals in the pedigree, which may require trimming the full pedigree into computationally tractable sub-pedigrees. The program PedCut (Liu *et al.* 2008) can construct a spanning set of sub-pedigrees that maximize the number of individuals who share a recent common ancestor given a maximum bit-size limit (bit-size is determined by the number of individuals and founders). IBD between individuals can be estimated with a Markov Chain Monte Carlo algorithm on the trimmed pedigrees, maintaining computational efficiency by using an informative (high minor allele frequency) set of SNPs in low linkage disequilibrium. Non-parametric multipoint linkage scans can be conducted by maximum likelihood estimation of variance components. The final step is the construction of a mixed linear model relating the sharing of genomic regions to phenotypic covariances and parameter estimation using restricted maximum

likelihood (George *et al.* 2000). The statistical significance of a possible QTL is obtained from the comparison of the likelihoods from a model with a putative QTL effect and one without. A number of programs have been developed for multipoint linkage analysis, such as Loki (Heath 1997) and SOLAR (Almasy and Blangero 1998). The program EMMAX (Kang *et al.* 2010) can perform finer scale mapping in regions near identified QTL using association tests that account for kinship.

1.2.2 Examine candidate regions for associations with fitness

Another approach to identifying fitness-related regions of the genome involves examining genes with ecologically significant functions. For instance, if disease were suspected to play a large role in causing mortality in a population, it would be useful to assess the operation of natural selection on genes involved in immune function using standard survival analysis. Correlations between candidate genes and survival can be tested using Cox proportional hazards models, accounting for the pedigree and including other possible explanatory variables (*e.g.*, year, sex, inbreeding coefficient) as covariates (Cox and Oakes 1984).

Combining these two approaches with the latest methods to scan genome-wide SNP data for regions under selection at deeper timescales (Sabeti *et al.* 2006; Grossman *et al.* 2010) will result in a collection of fitness-related markers. These can serve as foci for further study of recent selection by Selection Components Analysis. In addition, regions that show no signature of selection can form an initial set of putative neutral markers, which can be filtered based on commonly used criteria (*e.g.*, select SNPs in regions far from genes, pseudo-

genes, middle of introns, etc.; as in Andrés *et al.* 2010). These putative neutral regions can be used in further analysis of stochastic and demographic processes and as control regions in tests of inbreeding depression.

1.3 Characterize the dynamics of genome-wide variation in a declining population

Assaying genome-wide SNPs allows an examination of whether patterns of allele frequency dynamics are consistent across the genome, particularly focusing on the comparison of putatively neutral and fitness-related regions. This tests a major assumption of conservation genetics - that observed neutral genetic variation is a suitable proxy for accumulated detrimental variation and lost adaptive variation in small populations (Kohn *et al.* 2006). Considerable progress has been made in describing neutral variation in small populations, but the difficulty of identifying variation associated with selection has left unanswered questions regarding the dynamics of fitness-related variation (Angeloni *et al.* 2012).

If archived DNA samples exist, then it is possible to test whether recent population declines have resulted in a loss of genetic diversity. Mean heterozygosity and allele frequencies over time can be estimated from the genotype data. These observed values then can be compared to expected allele frequency changes from coalescent simulations of drift under two demographic scenarios (one with constant population size and one with decreasing population size) using recombination rates obtained from a linkage map.

Recent progress in modeling variation in a population where full genealog-

ical (pedigree) information is available provides more power to detect selection (Barton and Etheridge 2011), as the pedigree information allows far more sensitivity for departures from neutrality. Having genotype data for individuals on known pedigrees will give unprecedented ability to examine and test the role of random drift in changing allele frequencies. Rather than assuming that there is some particular sampling process for gametes, pedigrees allow direct observation of that sampling by comparing allele frequencies between parents and their progeny over time. Similarly, this full-pedigree approach will provide a powerful test of whether frequency dynamics are consistent across regions of the genome (allowing contrasts of neutral versus fitness-related regions). To determine whether neutral and fitness-related regions have similar dynamics, test whether a model incorporating both selection and drift is a better fit for observed dynamics at fitness-related markers compared to a pure drift model.

These approaches will not only determine whether genetic variation has been lost by drift over time, but also will empirically test the usefulness of neutral markers as a proxy for dynamics of fitness-related variation in a population. Comparing the patterns of neutral and fitness-related variation will inform future conservation genetics projects and provide additional insight into the balance between drift and selection.

1.4 Determine the role of selection in maintaining variation

Estimating the fitnesses of segregating genotypes in a population is notoriously difficult owing to the complex manner in which selection can act (Nadeau and Baccus 1981; Orr 2009). Measurement of net selection based on only changes in allele frequency measured at just a single stage of the life cycle is insufficient

(Prout 1965; Prout 1969). Rigorous inference of selection must consider all selection components: zygotic selection, sexual selection, fecundity selection, and gametic selection (Christiansen and Frydenberg 1973; Nadeau and Baccus 1981; Christiansen and Prout 2000). Evaluation of all four components requires detailed population monitoring, as observations need to be recorded at four different life-cycle stages (Christiansen and Frydenberg 1973). Comprehensive analyses of selection components relies on a tightly defined and hierarchical series of hypothesis tests on the various fitness components (Table 1.1). The Nadeau-Dietz-Tamarin selection component analysis (SCA) analyzes all progeny in a brood to test a hierarchy of independent hypotheses (outlined in Table 1.1) and has the ability to detect multiple selection components acting simultaneously (Nadeau *et al.* 1981).

For the SCA, individuals in the population at a given point in time are divided into five categories: mothers, mother-offspring pairs, progeny, nonbreeding females, and males. Table 1.1 describes the neutral expectations for genotype frequencies in the absence of specific selection components. For example, sexual selection in females is tested by comparing the genotypic frequencies in breeding and nonbreeding females in any given year. All other hypotheses are systematically tested in a similar manner. Detailed analysis of the selection components influencing different genotypes has the potential to provide an unparalleled understanding of the evolutionary processes operating in this population. This framework can be used to test for three possible mechanisms of balancing selection (see below).

Table 1.1: An outline of the selection component analysis (from Nadeau and Baccus 1981). Each hypothesis is tested by a χ^2 test statistic.

Component	**Null hypothesis**
Gametic selection (males)	For families with heterozygous fathers, both paternal alleles occur in equal frequency
Gametic selection (females)	Heterozygous females have 50% heterozygous offspring
Sexual selection (males)	Transmitted male gamete frequency equals genotypic frequency of adult males
Sexual selection (females)	Equal genotypic frequencies among mothers and nonbreeding females
Fecundity selection	Number of offspring similar for females of each genotype
Zygotic selection	Adult population same as estimated zygote population (all individuals have equal survival probabilities)
Random mating	Transmitted male gametes independent of genotype of the mother

It is also possible to expand the SCA framework, which relies on mother-offspring pairs, to include the full pedigree. For each genomic locus at each sampling point in time, the change in copy numbers of the observed alleles is constrained by the number of descendants each individual leaves. Patterns of observed copy numbers in the pedigree can be contrasted to null expectations generated from simulations dropping neutral variation on the pedigree (MacCluer *et al.* 1986), This simulation technique, called gene dropping, is a fast and powerful approach that simulates Mendelian transmission of alleles down a pedigree given founder genotypes (MacCluer *et al.* 1986). Sampling properties of the pedigree are discussed in Barton and Etheridge (2011). Briefly, the pedigree constrains the possible gene genealogies in a population. The genetic contribution of an individual can therefore be summarized by its reproductive value, which is the number of copies of its genes that are transmitted to future generations conditional on the pedigree (Barton and Etheridge 2011). A selectively favored mutation will be evident by the expansion in copy number of the haplotype bearing that mutation. This new approach of selection inference is more powerful than single generation methods, although it provides only inference of net selection effects because it integrates over the distinct selection components that SCA resolves.

1.4.1 Quantify temporal variation in selection pressures

Temporally-varying selection may play a role in maintaining polymorphism at a given genotype if the selection components operate differently across years (Haldane and Jayakar 1963). Given genotypes and detailed population monitoring from several different years, perform SCA on each year separately, then

perform tests of homogeneity through time. If the time homogeneity hypothesis is not rejected, then there is no evidence for variation in selection components among years for that genotype.

1.4.2 Identify polymorphisms experiencing age-specific variation in selection

Selective pressures can vary depending on age; hence the genetic architecture of fitness may change over an individual's life cycle. Studies have documented selection acting in opposite directions on a trait at different life history stages, and this phenomenon may contribute to the maintenance of genetic variation in natural populations (Schluter *et al.* 1991). The SCA framework can determine whether different selection pressures operate at different life stages on a given genotype. If detailed survival data exist, individuals can be sorted into different age classes for a finer analysis of viability selection.

1.4.3 Test the role of sexual conflict in the maintenance of polymorphism

Fitness-related loci may be selected in opposite directions in different sexes, leading to a trade-off that can affect both evolutionary trajectories and the maintenance of genetic variation (Foerster *et al.* 2007). Despite predictions of sexual conflict theory and laboratory experiments, little supporting evidence exists for sexually antagonistic genetic variation for fitness in wild populations (Foerster

et al. 2007). Positive evidence of sexually antagonistic variation in fitness can be obtained in two ways. (1) Performing linkage mapping for males and females separately can identify any QTLs that have opposite effects in the two sexes and negative genetic correlations between the two sexes. (2) Sex-specific differences in selection can be tested using SCA. In addition to parsing out sex-specific fecundity, sexual, and gametic selection components, one could analyze viability selection in males and females separately. This analysis has the potential to identify genotypes that experience different selection pressures in the two sexes. Finally, examination of the genomic distribution of sexually antagonistic loci can test the theoretical prediction that sexually antagonistic variation should accumulate on the Z chromosome (Rice 1984; Patten and Haig 2009).

Combining sensitive pedigree-based inferences of net selection with the fine-scale dissection of selection components using SCA can provide a thorough assessment of the role of selection in perturbing allele frequency dynamics in the population, and more specifically the role of selection in maintaining variation in the face of population decline.

1.5 Quantify the amount of new genetic variation introduced via gene flow

In small populations, a single immigrant can restore genetic variation originally lost to drift and inbreeding in small populations (Ingvarsson 2001; Hartl and Clark 2007). Individual-based population monitoring that can identify migrants and document immigration rate allows quantification of genetic migration rates each year and assessment of the degree to which they are associated with population declines as well as extrinsic factors such as weather patterns. With dense

genetic markers, one could trace the descendants of migrants and compare the genetic contribution of migrants to the population with that of resident birds on average or at different regions of the genome. Principal Components Analysis (Patterson *et al.* 2006) and STRUCTURE (Pritchard *et al.* 2000) can quantify the degree of genetic distinctness of the migrants from the residents. Given an assessment of genetic variants that appear to vary in fitness effects, gene dropping approaches (MacCluer *et al.* 1986) can determine the fitness impact of the novel migrant alleles. Fitting this information into a model for genetic variation can assess the net importance of migrants by comparing genetic diversity over time in models with versus without migrants. These approaches can test whether the contribution of migrants to population genetic variation has decreased over time.

1.6 Characterize the genetic architecture of inbreeding depression

Inbreeding is predicted to lead to decreased individual and population mean fitness (Kristensen *et al.* 2006). Characterizing the number and location of genes that affect inbreeding depression would help identify conditions that maximize purging efficiency (Allendorf *et al.* 2010). Gene expression differences between inbred and outbred lines of Drosophila suggest that inbreeding depression is associated with several genes, many of which are involved in metabolism, stress, and immune defense (Pedersen *et al.* 2005; Kristensen *et al.* 2006; Ayroles *et al.* 2009). However, these studies only sampled a small portion of the genome and primarily focused on male fitness (Ayroles *et al.* 2009; Paige 2010). In addition, since wild populations have different inbreeding and selection histories

than do lab-raised populations (Kruuk *et al.* 2000; Slate 2005; Pemberton 2008), and inbreeding depression can vary with environment (Slate 2005; Kruuk and Hill 2008), robust investigations of the genetic basis of inbreeding in the wild are essential for evaluating the conservation implications of population declines (Kristensen *et al.* 2010).

Testing for inbreeding avoidance (disassortative mating; a departure from random mating) can be done by contrasting genotypes of observed matings to expectations derived from computer randomizations. Combining detailed demographic data with genetic markers can yield a clear picture of the mechanisms by which inbreeding affects the reproductive fitness of individuals. The effect of inbreeding on measured phenotypic traits can be screened using animal models (as implemented in SOLAR). Traits for which the inbreeding coefficient is a predictor of the trait can be investigated in more detail. Genomic regions that may be contributing to these fitness effects can be identified using homozygosity mapping, which identifies regions of the genome that are homozygous in inbred individuals (Lander and Botstein 1987). These methods can identify genomic regions that contribute to inbreeding depression. Annotation of these regions can provide additional insight into the pathways involved in inbreeding depression and could clarify whether inbreeding depression is caused by increased homozygosity for recessive deleterious mutations or by overdominant loci.

1.7 A case study

The research questions described above can only be answered in declining organisms with well-described natural histories, long-term population monitor-

ing, and extensive genomic resources. One appropriate study system for testing the conservation genetics paradigm is a continuously studied population of the Florida Scrub-Jay (*Aphelocoma coerulescens*; hereafter FSJ), an iconic species on the U. S. Endangered Species List that has drastically decreased in number during the past half-century. The 43-year study has amassed detailed life histories and archived DNA samples for thousands of individuals, and the resulting 12-generation pedigree is one of the most accurate and extensive for any wild bird species (Figure 1.1; Clutton-Brock and Sheldon 2010). We use the FSJ as an example of the level of field and genomic resources required for a full characterization of the evolutionary processes associated with population decline.

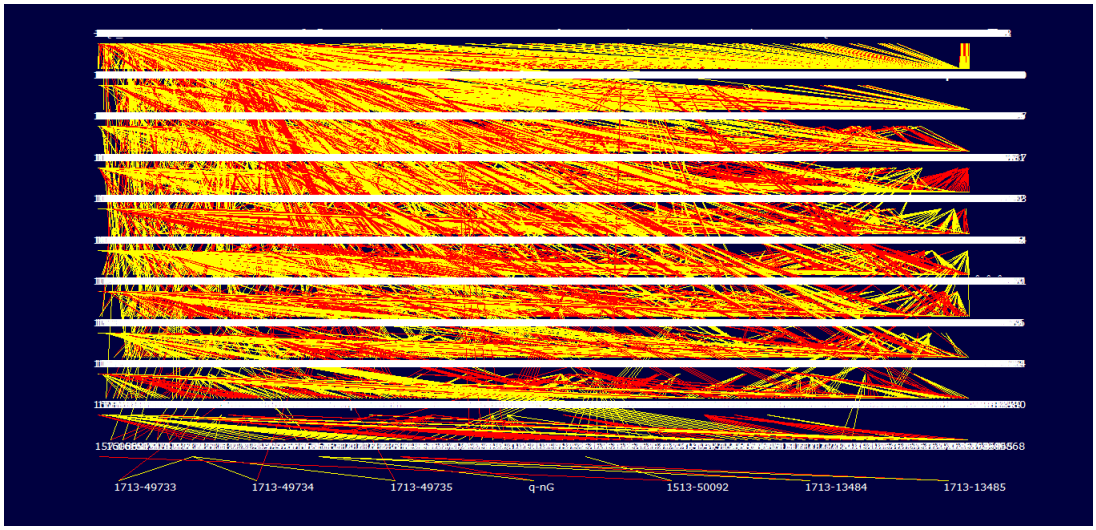

Figure 1.1: The full Archbold pedigree featuring 4,911 individuals over 12 generations. Lines show maternal (yellow) and paternal (red) relationships.

1.7.1 Natural history and demographic resources

The FSJ is the only bird species entirely endemic to Florida, and has one of the most narrowly restricted ranges in North America. Confined to fire-maintained, xeric oak scrubs that grow on well-drained sandy soils (*e.g.*, recent or ancient sand dunes), this non-migratory, cooperatively breeding species is thought to have once numbered at least 300,000 individuals range-wide (Woolfenden and Fitzpatrick 1996). Habitat loss from widespread agricultural conversion, residential and commercial development, and fire suppression has caused extensive population fragmentation and local extirpation. A thorough survey in 1992-93 suggested that the population had declined 97% in 100 years and 25% in the previous ten years alone. At that time, the total population size was composed of about 10,000 individuals living in ~3,000 family groups (Fitzpatrick *et al.* 1994). The newest statewide survey, completed in 2010, documented further decline: the species is now down to ~6,000 individuals in 2,400 family groups (Boughton and Bowman 2011). Although common SNPs that were segregating in the initial population are still likely to be segregating in the reduced population today, this magnitude of reduction has inevitably resulted in a severe loss of rare alleles and of haplotypes.

Intensive study of a FSJ population located at Archbold Biological Station, near the southern tip of the Lake Wales Ridge (LWR), began in 1969. Significant tracts of the scrub ecosystem in the vicinity of Archbold have been protected from development and are managed for conservation (Turner *et al.* 2006). The FSJ population in this immediate area remains stable (~200 family groups are fully protected by Archbold and surrounding state-owned preserves), but the surrounding LWR region continues to experience dramatic declines in both

habitat and jays as the human population increases (FSJ numbers in the LWR region have declined 85-90% in just 50 years; Boughton and Bowman 2011). Simulations suggest that a sustained 97% reduction in population size would reduce species-wide genetic diversity by 97%; reduced influx of novel alleles into Archbold would result in a similar loss of diversity within a few generations. Moreover, effective dispersal of FSJs decreases as habitat fragmentation increases (Coulon *et al.* 2010), further eroding opportunity for the arrival of new alleles. Thus, while the local FSJ population at Archbold has been stable, genetic erosion is likely because regional declines have isolated the population and reduced the influx of genetic diversity by immigration. Despite increased habitat conservation efforts (mainly land acquisition and prescribed burning), individual-based population monitoring since 1969 found that immigration into Archbold has decreased over time since 1990 (Figure 1.2; the apparent increase in the 1980s is likely due to expansion of good habitat at Archbold caused by initiation of prescribed burning regimes). Well-documented dispersal curves based on banding studies (Coulon *et al.* 2010) suggest that almost all unknown-origin immigrants into the study area are from nearby territories; numerous lines of evidence indicate that the dramatic drop in immigration rate reflects diminished influx from more distant habitat sources, a result that has been exactly mirrored in a nearby suburban study area (Bowman, unpubl. data).

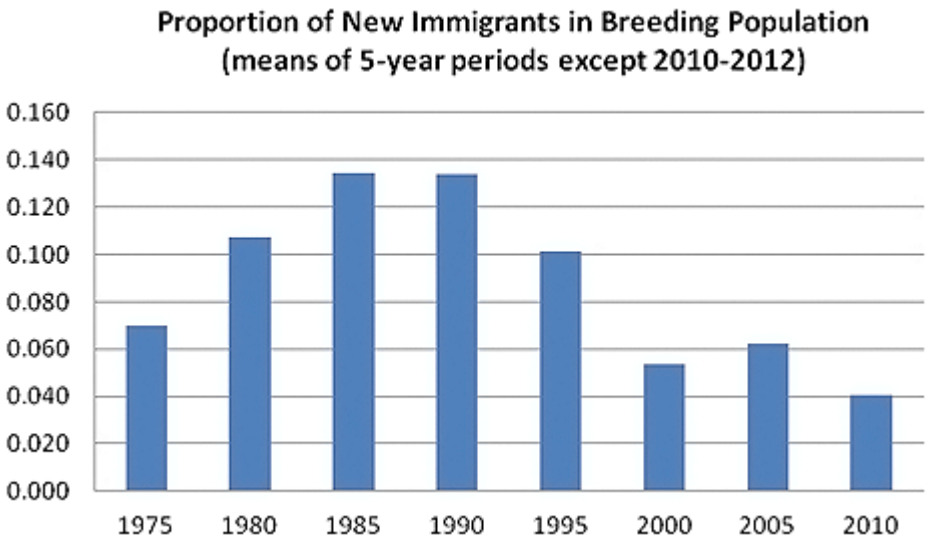

Figure 1.2: Mean proportion of breeders that are unknown-origin immigrants, in five-year increments. Immigration into Archbold has decreased over the past 20 years.

This study population of FSJs provides an unparalleled model for research on the genetics of fitness in the wild, making the FSJ a prime candidate for a conservation model species. Since 1969, every nestling and immigrant has been uniquely banded, and a complete census is conducted each month. The FSJ is highly sedentary, so a large proportion of surviving offspring breed within our 80-territory study area. Surviving offspring remain in the family group for at least one year before leaving home, and all adults remain attached to a neighborhood once they become breeders. As a result, mortality of both juveniles and breeders is documented directly and precisely by monthly censuses, providing exceptionally accurate data on individual lifespans (Woolfenden and Fitzpatrick 1984; Woolfenden and Fitzpatrick 1996). FSJs are genetically monogamous (Quinn *et al.* 1999; Townsend *et al.* 2011); therefore field observations alone provide all information necessary to construct accurate pedigrees. Every nest within every family-group is monitored (for clutch sizes, nestlings, and fledglings), producing fully documented annual fecundity and lifetime fitness measures for all birds of known age (Table 1.2). Morphological traits are recorded for each individual at each key stage (nestling, independent, adult; Table 1.2). From 1988 to 1995 and every year since 1999, blood is sampled from every nestling and immigrant recruited into the population. Blood samples and high-quality extracted DNA aliquots are stored in separate archives for 4,017 jays with known-parentage and full demographic history. The largest pedigree includes 4,420 individuals and is 12 generations deep (Figure 1.1). Relevant ecological data include annual habitat composition of every FSJ territory, daily weather records, annual or monthly measures of food availability, monthly indices of predator abundance, and detailed spatial history of fires.

Table 1.2: List of measured phenotypes in the Archbold population. Viral antibody titres were only collected from 2007 onwards.

Category	Phenotype	Longitudinal measure
Life history	Lifespan, Annual fecundity, Lifetime fecundity (#eggs through to #grand children), Age at natal dispersal, Age at first breeding, Breeding lifespan, Nest phenology, Nest site attributes, Habitat preference, Clutch size	
Morphometrics	Tarsus length, Tail (6th retrix) length, Wing (7th primary) length, Bill length/depth/width	Day 11, Day 85, subsequent captures
Body condition	Mass, Viral antibody titres	Day 11, Day 85, subsequent captures
	Fat content, Ectoparasite load, Microfilaria load	Day 85, subsequent captures

Realized lifetime fitness varies widely in the study population, and has been shown to be heritable (Fox *et al.* 2006; Chen and Van Hout unpubl. data). Only about 35% of fledglings survive to age 1, and of those that become breeders, lifetime reproductive success in both sexes is correlated with lifespan (Figure 1.3; Woolfenden and Fitzpatrick 1991). Natural environmental elements affecting individual survival include a suite of vertebrate predators, four arboviruses, two protozoan blood parasites, and an array of endoparasites (Kinsella 1974; Garvin *et al.* 2004), including microfilaria, which correlates with juvenile survival (Robbins and Boughton unpubl. data).

The FSJ is one of the longest-studied endangered species. This wealth of environmental, demographic, and morphometric data from a population with known genealogy is crucial for thorough evaluation of basic population and conservation genetics questions regarding the genetic basis of fitness variation and population genetic responses to declining population size.

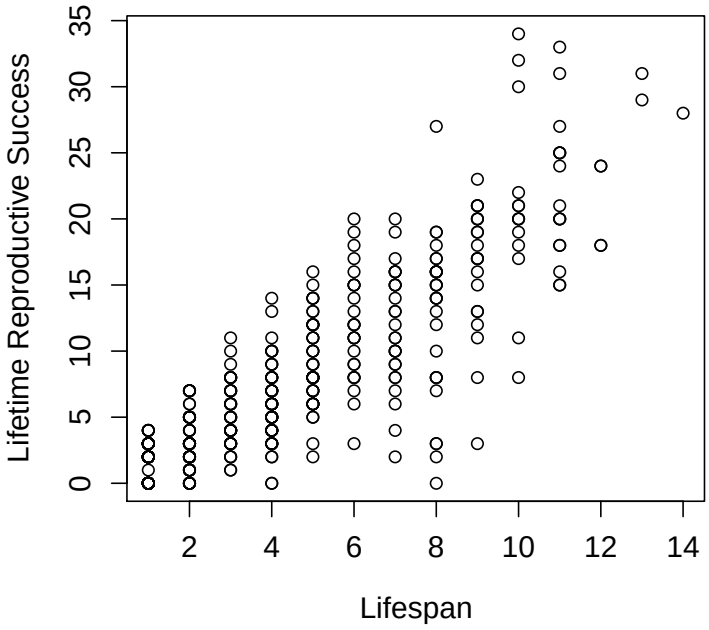

Figure 1.3: Lifetime reproductive success, as measured by total number of fledglings produced, is correlated with lifespan.

1.7.2 Genomic resources

Significant genomic resources have been developed for the FSJ. The genome of an inbred male FSJ has been sequenced to 107x coverage, and the preliminary assembly has 99,631 contigs and 14,020 scaffolds, with N50s of 24.2 kb and 264 kb respectively. Transcriptomes of a panel of diverse tissue samples from one male and one female have been sequenced by random-primed RNA-seq. A *de novo* assembly of these reads using Trinity (Grabherr *et al.* 2011) resulted in 32,394 unique transcripts, with an average length of 2.5 kb, and 31% of these reflect unique one-to-one orthology with transcripts of the chicken. Annotation of the FSJ genome should be reasonably robust given the high degree of synteny among extant bird lineages (which increases the usefulness of other annotated bird genomes; Ellegren 2010) and the transcriptome data. Even these preliminary assemblies are sufficient to greatly inform the analyses described above.

A genome-wide set of SNPs in the FSJ population was generated by using a modified protocol for genotyping-by-sequencing (Elshire *et al.* 2011) to sequence 107 individuals in 28 families. Spurious SNPs were filtered with a probabilistic, reference-free pipeline that takes advantage of known family relationships among individuals to perform statistical tests of consistency with Mendelian inheritance (Chen *et al.* 2014). Illumina custom iSelect BeadChips (http://www.illumina.com/applications.ilmn) were designed for a set of 15,000 high quality genome-wide survey SNPs, and all individuals in the sample archive (from 1988-2012) were genotyped. This dense genotyping is expected to be sufficient for measuring recombination with high resolution based on average recombination rate estimates in Zebra Finch. Genotype data combined with available pedigree and demographic data permit a number of differ-

ent analyses investigating the population genetic processes affecting patterns of neutral and fitness-related variation.

1.8 Conclusions

We now have the potential to combine genomics with theoretical population genetics, conservation genetics, and exceptionally well-studied field biology systems. The application of genomics to long-term demographic studies allows a comprehensive investigation of patterns and processes affecting genetic variation at a genome-wide scale in the wild. It is axiomatic in conservation that preserving genetic variation is important, but with modern genomic technologies we finally can test the validity of this assumption. Specifically, testing whether neutral marker variation presents an accurate proxy for variation at functionally important regions of the genome is crucial for informing future conservation efforts and policies. We provided guidelines on how to identify fitness-related regions of the genome and investigate whether genetic variation at different regions of the genome has been lost over time due to drift and reduced gene flow. Conservation genomic studies can also investigate the maintenance of variation by selection and characterize the genetic architecture of inbreeding depression. Each of the described approaches focuses on a particular evolutionary process that influences population genetics. The larger picture of how genetic variation is governed in a population involves the interaction of all these forces (plus novel mutations). Incorporating the individual parameter estimates into one cohesive model will present an unparalleled understanding of the dynamics of genetic variation in a real-world setting. Achieving such a model is especially vital in the context of conserving species undergoing range-wide fragmentation and decline.

CHAPTER 2

IDENTIFICATION OF AVIAN W-LINKED CONTIGS BY SHORT-READ SEQUENCING

2.1 Abstract

The female-specific W chromosomes and male-specific Y chromosomes have proven difficult to assemble with whole-genome shotgun methods, creating a demand for new approaches to identify sequence contigs specific to these sex chromosomes. Here, we develop and apply a novel method for identifying sequences that are W-specific. Using the Illumina Genome Analyzer, we generated sequence reads from a male domestic chicken (ZZ) and mapped them to the existing female (ZW) genome sequence. This method allowed us to identify segments of the female genome that are underrepresented in the male genome and are therefore likely to be female specific. We developed a Bayesian classifier to automate the calling of W-linked contigs and successfully identified more than 60 novel W-specific sequences. Our classifier can be applied to improve heterogametic whole-genome shotgun assemblies of the W or Y chromosome of any organism. This study greatly improves our knowledge of the W chromosome and will enhance future studies of avian sex determination and sex chromosome evolution.

2.2 Background

While whole-genome shotgun and short-read assemblies are rather effective at reconstructing single-copy euchromatic genes, repetitive regions remain a major challenge. Short-read sequencing eliminates issues related to low cloning

efficiency of interspersed repeats, but the assembly process remains problematic for both repeats and segmental duplications, as high sequence homogeneity among copies of a given repeat or duplication limit the potential to reconstruct sequence order (Mardis 2008; Alkan *et al.* 2011). The inability to assemble repetitive regions can also pose difficulties for reconstructing large scaffolds from contigs (Green 2001), and the resulting gene fragmentation complicates gene assembly and annotation (Alkan *et al.* 2011). The assembly of repeats and duplications therefore remains a major challenge in genome sequencing and is only possible by focused and concerted efforts (Schueler *et al.* 2001; Skaletsky *et al.* 2003).

In species with chromosomal sex determination, the male-specific Y (in species with XX/XY sex determination) and female-specific W chromosomes (in species with ZZ/ZW sex determination) present special challenges to whole genome shotgun assembly. Sex-specific chromosomes are enriched for interspersed repeats and segmental duplications, on which whole genome shotgun methods perform poorly (Skaletsky *et al.* 2003). The absence of crossing-over outside the pseudoautosomal region makes it impossible to take advantage of the genetic map for scaffolding the assembly (Foote *et al.* 1992). An additional hindrance is the lower sequence coverage of the sex chromosomes when sequencing heterogametic individuals, which reduces the average length of assembled contigs. Sex chromosomes receive half the coverage of autosomes when sequencing heterogametic individuals (the strategy used for chicken and turkey), and just a quarter of the autosomal coverage if sequencing a 50:50 mix of heterogametic and homogametic individuals (the strategy adopted for *Drosophila melanogaster*). Even in organisms like *Drosophila melanogaster*, where the quality of the whole genome shotgun assembly is extremely high, the Y

chromosome remains a collection of unassembled contigs (Hoskins *et al.* 2002; Carvalho *et al.* 2003; Hoskins *et al.* 2007). In the case of humans and chimpanzee, the Y chromosome assemblies are nearly complete, because these were sequenced by a painstaking BAC-by-BAC effort (Skaletsky *et al.* 2003; Hughes *et al.* 2010).

There is considerable interest in assembling the female-specific avian W chromosome, not only to expand our understanding of sex-determination mechanisms, but also to address many questions about sex chromosome evolution. The exact mechanism of avian sex determination remains controversial: though the Z-linked *DMRT1* gene is required for testis development (which is consistent with the Z dosage hypothesis), female sex determination may still involve a dominant, W-linked gene (analogous to Y-linked dominant sex determination in mammals; Ellegren 2000; Smith *et al.* 2009b). More information about the W chromosome will contribute to our understanding of the evolution of female heterogamety as well as the dynamics of sex chromosome degradation and differentiation (Mank and Ellegren 2007).

The chicken genome, which contains 38 autosomes and a pair of sex chromosomes, was sequenced in 2004 from a single female Red Junglefowl (International Chicken Genome Sequencing Consortium 2004b). About 70% of the heterochromatic chicken W chromosome consists of *Xho*I-, *Eco*RI-, and *Ssp*I-family repetitive sequences, and some known genes on the W are tandemly duplicated (*e.g.*, *Wpkci* Hori *et al.* 2000), leaving an estimated 10-15 Mb of non-redundant sequence (Itoh and Mizuno 2002). The chicken genome was sequenced to 6.6x coverage and assembled from whole-genome shotgun reads, as well as plasmid, fosmid, and bacterial artificial chromosome (BAC)-end read pairs (International

Chicken Genome Sequencing Consortium 2004b). Of the 1.05 Gb of assembled sequence, only 933 Mb were anchored to a specific chromosome, leaving 121 Mb in unmapped sequence fragments, collectively called chrUn (International Chicken Genome Sequencing Consortium 2004b). Assembly of the W chromosome is especially poor: only 0.5% of the W (based on its estimated size of 50-55 Mb) has been successfully mapped. To date, only a handful of genes have been identified on the W: *CHD1W* (Ellegren 1996), *ATP5A1W* (Fridolfsson *et al.* 1998), *ASW/Wpkci/HINT1W* (Hori *et al.* 2000; O'Neill *et al.* 2000), *SPINW* (Itoh *et al.* 2001), *SMAD2* (Itoh and Mizuno 2002), *UBAP2W/ADO12W* (Axelsson *et al.* 2004), *ZNF532W* (Wahlberg *et al.* 2007), *ZFRW* (Wahlberg *et al.* 2007), *MIER3W, hnRNPKW* (Nam and Ellegren 2008), *SSC2W/NIPBLW*, and *KCMFW* (first identified in Build 2.1 and then cited by Nam and Ellegren 2008).

Given the challenges in producing an assembly of the Y and W chromosomes by traditional shotgun-sequencing methods, new tools are required to identify sex-specific sequences generated by heterogametic shotgun sequencing projects. Here, we adapt a method devised by Carvalho and colleagues (Carvalho and Clark 2013) and identify female-specific sequences by contrasting male-derived, short-read shotgun genomic sequences and unmapped sequence fragments (chrUn) from the female-derived chicken genome. This method relies on the fact that the W chromosome is female-limited. By sequencing the genome of the homogametic sex (in our case, the ZZ male) to high depth and aligning the reads to the genome of the heterogametic sex (the ZW female), we were able to identify regions of the genome that are underrepresented in males and are therefore likely to be female-specific.

2.3 Results

2.3.1 Conceptual framework

Because avian males carry two Z chromosomes, the male genome should not contain any sequence that is found exclusively on the W chromosome. Thus, when mapping reads generated from a male (ZZ) back to the shotgun genome assembly generated from a female (ZW), very few, if any, reads should uniquely map to segments of the female ZW genome that are W-specific. In particular, evidence that unmapped contigs from the ZW female are likely to be W-specific derives from their under-recruitment of matches to the reads from ZZ males (see overview of method in Figure 3.1). This method is similar to the read depth approaches for detecting copy number variants, which assume a Poisson distribution in mapping depth and therefore detect duplications and deletions by searching for regions with significantly higher or lower read depth (Medvedev *et al.* 2009; Alkan *et al.* 2011). Our pipeline relies on the subtraction of the male genome from the female genome and tests for lower read depth on a contig-by-contig basis. We summarize alignment results for each contig using both the number of unmasked bases covered by a read (coverage) and a normalized measure of total number of reads aligned (read depth; Figure 3.1B). Both measures should be near zero for W-specific contigs but not for autosomal or Z-linked contigs (Figure 3.1C).

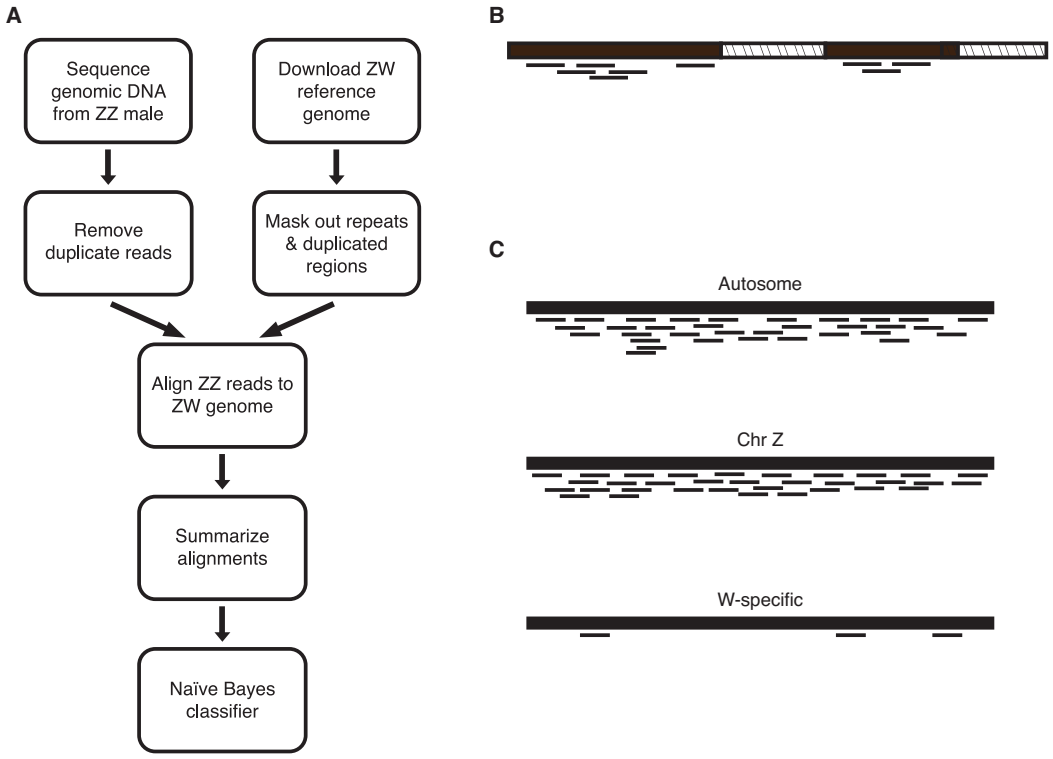

Figure 2.1: A novel method of identifying W-specific contigs. (A) The steps in our classification procedure. **(B)** Alignment results were summarized by two statistics: coverage and read depth. If a contig consists of unique sequence (solid black) and masked repetitive regions (hatched), then coverage is the proportion of unique sequence covered by a read. Read depth is the number of reads divided by the total possible locations to which a read could map. **(C)** Predicted alignment results. Each W-specific contig should have very few male-derived reads uniquely aligning to it.

2.3.2 W-specific contigs have distinct coverages and read depths

We generated roughly 10 million reads from a ZZ individual and mapped them to the unique regions of the ZW genome. As predicted, the previously-mapped W-linked contigs had significantly fewer uniquely aligned reads relative to known autosomal and Z-linked contigs (Figure 3.2). The known W-contigs have coverage and read depth values near zero: 95% bootstrap CI for coverage is (0, 0.083) and for read depth ($2.24x10^{-6}$, 0.0139). This was expected because sequences derived from a male genome are not expected to map to W-linked contigs. In contrast, male-derived sequences readily align to known autosomal and Z-linked contigs. Autosomal/Z-linked contigs have non-zero read depths and coverages: the 95% bootstrap CI for coverage (0.256, 0.293) and for read depth (0.00862, 0.00992) both are positive. Thus W contigs have significantly different coverage and read depth values than autosomal/Z-linked contigs ($p = 0$).

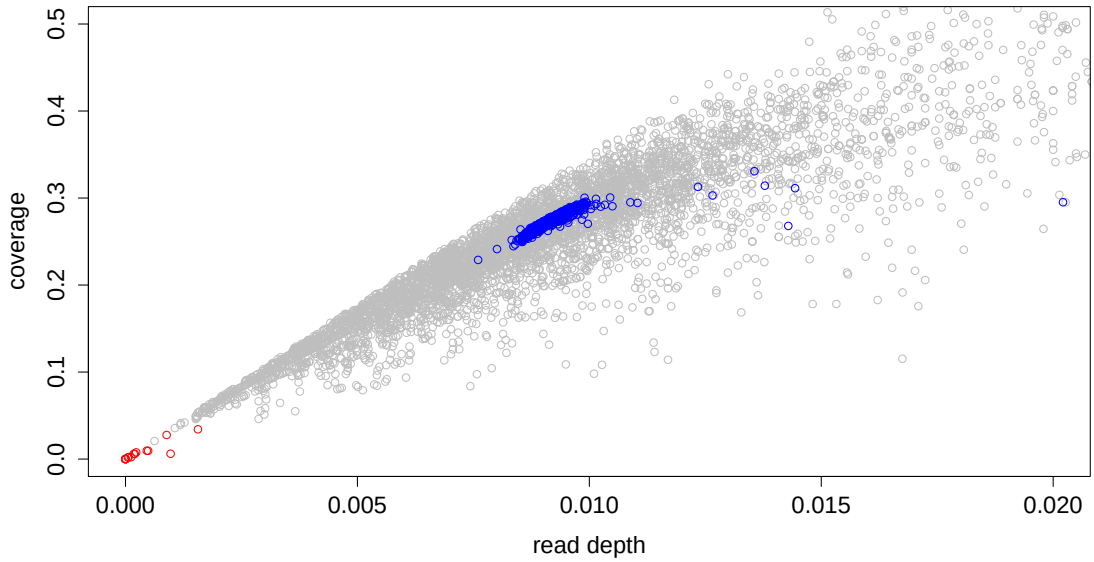

Figure 2.2: Discrimination between W and non-W contigs based on read depth and coverage for each contig. Known W contigs are shown in red, known Z or autosomal contigs are in blue, and unmapped contigs are in gray. Note that the W contigs (red dots) exhibit very low alignment and read depth to male-derived sequences. The W contigs form a distinct cluster from the autosomal or Z contigs. The goal is to classify all the unmapped contigs (gray dots) into one of two classes: W or non-W.

2.3.3 Contig length influences alignment results

Due to the stochasticity of the sequencing method, the length of the contig may affect the distribution of hits along the contig and therefore our prior expectations of both coverage and read depth. We simulated several genomes, each with contigs of a different length. Once contig length decreased to 1,500 bp or less, the probability that an autosomal or Z-linked contig would be misclassified as a W-specific contig increased exponentially (Figure 3.3). After stringent filtering, 57% of the remaining 6,905 unmapped contigs are of length 1,500 bp or less. It is therefore important to take contig length into consideration in the classification method. Not accounting for the fact that very short contigs have fewer hits regardless of class would greatly inflate the false positive rate of the classification approach.

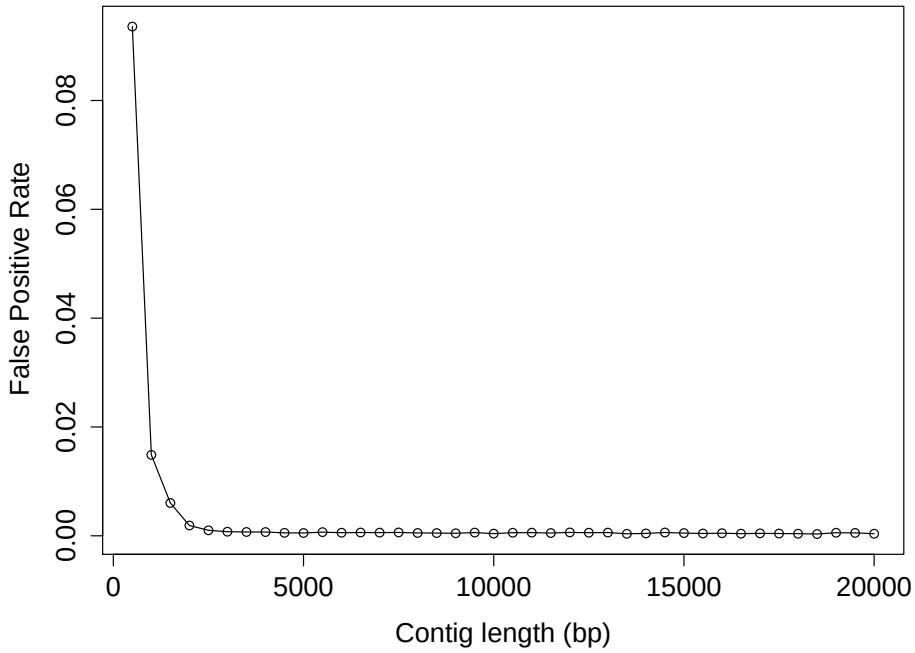

Figure 2.3: False positive rate as a function of contig length. Here the false positive rate refers to the fraction of known autosomal or Z-linked contigs with coverage less than 0.10 (the mean 100th quantile for coverage of known W-specific contigs from the bootstrap replicates).

2.3.4 Evaluation of performance

We developed a naïve Bayes classifier to determine which of the unmapped contigs are likely to be W-specific. A naïve Bayes classifier relies on a set of training data to estimate parameters for classification. Thus properties of the training set may significantly affect the performance of the classifier. We performed several different experiments to optimize the classifier. By running cross-validation tests with the previously mapped contigs, we investigated the effects of training set size, sample imbalance, and bin sizes of the feature distributions on the classifier. ROC curves were generated and the area under the curve calculated for all variations of the method. Increased training set size improved the performance of the classifier (Figure 3.4). This result is not surprising: the more data used to estimate model parameters, the better the classifier performs. Sample imbalance occurs when there is unequal representation of different classes in a dataset. Imbalanced datasets can negatively impact the performance of machine learning algorithms. However, in our case, sample imbalance did not seem to be a problem: we ran the classifier with different ratios of non-W:W contigs (from 1:1 to 100:1) in the training set and found no significant differences in performance. Finally, we also tested different variations of the feature probability distributions. Evaluation of the different bin sizes for discretizing distributions of coverage and read depth found that the optimal bin size is 0.005.

After optimizing the classifier using known data, the next step was to evaluate the ability of the classifier to accurately predict novel W sequences. Our classifier identified 629 candidate W-specific contigs from the set of unmapped contigs. We have tested 315 contigs by PCR and confirmed 62 of them as female-specific (Table A.1). Of these, we found female-specific markers on 51 of the 177

contigs that had a >95% posterior probability of being W-specific. We used these results to further evaluate the sensitivity and specificity of our method in independent data set tests. In these tests, the contigs of known location were used to train the data set, and performance of the classifier was evaluated using the PCR-confirmed set. A series of independent data set tests were used to test the effects of contig length and sample imbalance on classifier performance. Our simulations (see above) predicted that contig length should influence classification results. To test this prediction, we used contigs of varying sizes to train the classifier and compared performance on the same validation set of short (mostly 1 kb) contigs. Classifier performance decreased substantially when >10 kb contigs were used to train the classifier. Contig length does affect classification results, which explains why greater accuracy is achieved by conditioning on contig length (Figure 3.5A). Unlike the results from our cross-validation tests, sample imbalance had more of an effect in these independent data set tests. Performance improves slightly when the non-W:W ratio is below 10 (Figure 3.5B); therefore, severely unequal representation of the non-W and W classes affects the predictive performance of the classifier. However, sampling methods such as over-sampling the minority class (W) or under-sampling the majority class (non-W) can achieve better results. Overall, the classifier did not perform as well in the independent data set tests, most likely due to the high false positive rate that resulted from insufficient sequence coverage.

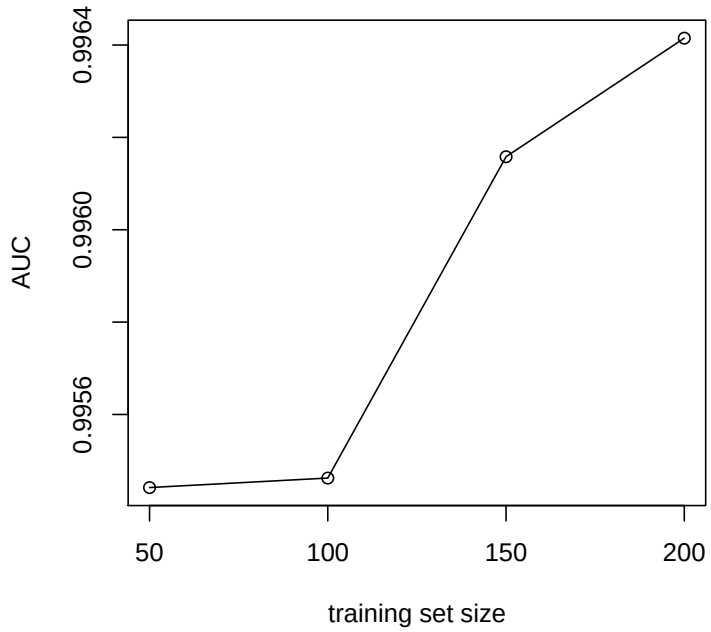

Figure 2.4: Performance of the classifier as a function of number of contigs in the training set. We ran the classifier with increasing numbers of contigs, from 50 W and 50 non-W contigs to 200 W and 200 non-W contigs. This was done by subsetting the mapped contigs 100 times: for each iteration, the set of training contigs was randomly selected, and the remainder used for validation. The mean AUC for each training set size is shown. AUC (area under the ROC curve) is a commonly used statistic for model comparison.

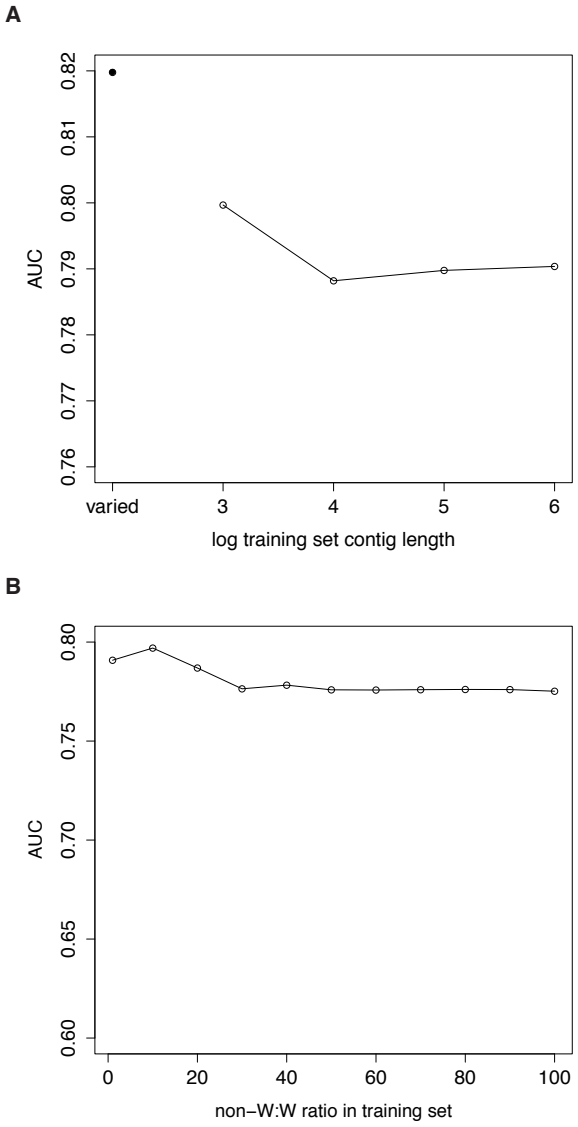

Figure 2.5: Performance of the classifier as a function of contig length and training set composition. Here the validation set consists of the confirmed chrUn contigs, and the training set is a subset of the set of mapped contigs. **(A)** Contig length matters. The open circles show results without conditioning on length; instead, the same validation set was classified using training sets with different contig lengths (1 kb - 1000 kb). For each contig length, we randomly selected 200 W and 200 non-W contigs for the training set. This was performed 100 times. The validation set contigs are short (average <1 kb in length), and the classifier performs better when shorter contigs are used for training. However, performance is maximized when we condition on length in the classifier (solid circle). Classifier performance is measured by mean AUC. **(B)** AUC for different ratios of non-W to W contigs in the training set. AUC increases for smaller non-W:W ratios.

2.4 Discussion

We present a framework to identify W-specific sequences in the chicken genome. The approach is generalizable to identify any genomic sequences that are present uniquely in one sex (*e.g.*, Y or W chromosomes within other animal species), and is potentially useful for characterizing the genomes of non-model organisms. Our method is based on the fact that sequences unique to the W chromosome are not present in the genome of a male. We mapped male-derived sequence fragments to the genome of a female and developed a naïve Bayes classifier using the alignment results (summarized by coverage and read depth). As predicted, contigs specific to the W chromosome had significantly lower coverages and read depths.

The accuracy of our method can be improved with deeper sequencing. Many of the false positive contigs probably had low coverages and read depths due to low sequence depth. We generated 367.2 Mbp of high quality sequence, which translates to only 0.45x coverage of the masked genome. It is therefore not surprising to find portions of the genome misleadingly underrepresented in the data set. At half this coverage, 40% of contigs of length 1 kb have very few reads aligning, making it more difficult to distinguish true female-specific contigs. However, this depth of sequencing was sufficient for proof of concept. We show that, even at low coverage, the approach was successful at identifying a focal set of candidate sequences for subsequent verification by targeted PCR.

Unlike traditional sequence mapping methods, our approach is not severely hindered by the lower sequence coverage of the W chromosome during shotgun sequencing of heterogametic individuals. While lower coverage results in W contigs that on average are shorter in length (and therefore more difficult

to classify), we greatly improve performance by conditioning on contig length in the classification method. However, our method cannot fully overcome the challenges posed by repetitive regions. All interspersed repeats and segmental duplications were masked out of the genome before performing the alignments, thereby eliminating much of the W chromosome from consideration. It is possible to relax the stringency of the filtering step in further iterations of the classifier to identify euchromatic repeats that do not resemble genome typical repeats. Furthermore, this method cannot exhaustively find all non-repetitive W contigs - it can only detect unique regions specific to the W. Sequences in the pseudoautosomal region will produce the same read depth as autosomal regions, and recent gene duplication events may produce W-linked sequences with enough similarity to autosomal or Z-linked sequence to be represented in male genomes.

Because our method searches for regions in the male genome that are underrepresented in female-derived genome sequences, any male-specific deletions could lead to an inappropriate assignment of contigs to the W chromosome. Deletions in the White Leghorn genome compared to the Red Junglefowl genome are not an issue because all our PCR validations used males and females of the same species. Our method would classify a deletion in the White Leghorn genome as W-specific, but such a region would not show a female-specific amplification pattern in our PCR validation step. Misclassifications due to male-specific deletions can be detected by screening a larger set of individuals and by BAC screening and sequencing.

Despite the limitations of our approach, we were still able to identify more than 62 new W-specific contigs. Note that this number is an underestimate, as

contigs that fail to produce a female-specific marker may still be located on the W chromosome. These new markers will greatly improve the assembly and annotation of the W chromosome. A more complete annotation of genes on the chicken W chromosome will accompany the BAC-based sequencing and assembly of the chromosome.

There is particular interest in fully annotating the avian W because the sex-determining mechanism in birds has yet to be completely characterized. *DMRT1* is known to be required for testis development (Smith *et al.* 2009b), though studies on triploid and chimeric chickens suggest there may be a female-determining gene that interacts with a male-determining locus on the Z (Smith and Sinclair 2004; Smith *et al.* 2009a). Evidence supporting the popular W-linked candidate, *HINTW*, is mixed: though *HINTW* is functionally different from its Z chromosome paralog (Hori *et al.* 2000), mis-expression of *HINTW* in male (ZZ) embryos resulted in normal testes development (Smith *et al.* 2009a). Further annotation of the W may unearth other candidate ovary-determining genes.

Sequence information of the W chromosome would benefit several different evolutionary studies besides avian sex determination, from sex chromosome evolution to sexual conflict and sex-biased mutation rate (Ellegren 2000). For example, birds are good subjects for the study of sex chromosome evolution because different bird groups exhibit parallel divergence of the W as well as variation in the degree of W chromosome degradation (from a largely undifferentiated state in ratites to a highly degenerate state in passerines; Shetty *et al.* 1999; Mank and Ellegren 2007). The scope for genetic conflict is increased in ZW species because the W is expressed in both sexes in the form of maternal effects, and the accumulation of sexually antagonistic maternal effect genes could

contribute to the decay of the non-recombining W (Miller *et al.* 2006). The W chromosome may be a magnet for female-specific fertility genes. Evolutionary theory indicates that male fertility genes are expected to be retained on the Y chromosome because they are free from the influence of selection in females (Fisher 1931; Roldan and Gomendio 1999). By symmetry, this same evolutionary theory leads to the expectation that the W chromosome may concentrate genes that are uniquely necessary for female fertility (Fisher 1931; Roldan and Gomendio 1999). Finally, ZW systems may be more appropriate than XY systems for studying sex-specific mutation rates: while higher mutation on the Y may be due to male-biased mutation or suppressed mutation on the X chromosome to minimize exposure of deleterious recessives in the hemizygote male, these hypotheses can be distinguished in ZW sex chromosomes (Ellegren 2007).

The availability of more W-specific sequences also facilitates the development of additional sex-specific primers for unambiguous molecular sexing techniques. The ability to sex individuals is critical for answering several questions in evolution and ecology, and morphological identification of sex is often difficult in birds (Ellegren and Sheldon 1997). The commonly used universal primer sets for avian molecular sexing depend on length differences between *CHD-Z* and *CHD-W* introns (Griffiths *et al.* 1998; Kahn *et al.* 1998; Fridolfsson and Ellegren 1999), which may be problematic in certain species due to *CHD-Z* polymorphisms (Dawson *et al.* 2001) and heteroduplex molecule formation (Casey *et al.* 2009). Thus the new W-specific sequences identified here can help advance several different avenues of research.

2.5 Conclusions

Here we describe a novel approach for identifying sequences specific to a heterogametic sex chromosome. We performed a proof-of-concept experiment by aligning shotgun sequence reads from a male (ZZ) chicken to the genome of a female (ZW) chicken, and our classifier successfully identified >60 confirmed novel W-specific contigs despite low coverage. We believe that our method is widely applicable and can benefit future genome assembly efforts. While there have been significant investments in lowering sequencing costs and increasing sequencing throughput, little investment has been made in techniques to cope with the limitations of whole-genome shotgun sequencing strategies, particularly the challenges specific to sex chromosomes: low coverage, resolution of interspersed repeats and segmental duplications, inability to map, etc. In addition, *de novo* assemblies generated using only next-generation sequencing technologies are especially prone to collapsing segmental duplications and large repeats (Alkan *et al.* 2011). The approach described here can quickly identify candidate W or Y chromosome markers, and these contigs can be extended by probing BAC libraries. A full assembly of the W chromosome still requires substantial BAC sequencing efforts, but this method can greatly facilitate the process of designing W-specific probes. A combination of our method with traditional BAC screening and sequencing would provide a powerful approach to assembling the W or Y chromosome in any organism.

2.6 Methods

2.6.1 Data generation

Genomic DNA was extracted from the blood of a White Leghorn rooster using the Qiagen DNeasy kit. We generated 10.5 million 36 bp reads using the Illumina Genome Analyzer (GA-IIx). Duplicate and low-complexity reads were removed before alignment, resulting in a total of 10.2 million unique and high quality reads. The sequence data generated in this study have been submitted to the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession SRP008449. We obtained chicken genome sequences (Build 2.1) and known W chromosome BAC sequences. The chicken genome assembly includes 18 scaffolds mapped to the W chromosome, and 1,044 autosomal or Z-linked scaffolds. The 25,378 unmapped contigs (chrUn) had lengths ranging from 54 to 48,370 bp. Low complexity sequences and repeats were masked with RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker/). After removing segments less than 50 bp in length, this resulted in 920.7 Mbp of sequence and 20,069 unmapped contigs. However, because our method relies on the unique mapping of reads, any sequences that occur in multiple locations in the genome could lead to spurious results. Thus, more stringent filtering of the reference genome was required. We aligned the masked contigs to themselves in MUMMER (Kurtz *et al.* 2004) and masked any duplicate regions larger than 50 bp. After this more stringent filtering step, we were left with a total 823.7 Mbp of unique sequence, with 6,905 unmapped contigs. Reads were aligned to the masked and filtered reference genome using MAQ (Li *et al.* 2008). We allowed some mismatches in the alignment process to account for sequence divergence between White Leghorn and Red Junglefowl (International Chicken

Genome Sequencing Consortium 2004a). Alignment results were summarized for each contig using two statistics: coverage and read depth (Figure 3.1B). Here we define coverage as the fraction of unmasked bases in a contig that is covered by one or more reads. Read depth is the number of reads aligning to a contig, normalized by the total number of locations a read could align to that contig. Our measure of read depth is analogous to the widely used measure of gene expression, reads per kilobase of exon model per million mapped reads (RPKM). Because we used only one library, there was no reason to calculate RPKM, which standardizes among libraries.

2.6.2 Confirmation of predictions

Because a large portion of the initial chicken W chromosome assembly was later discovered to be misassigned (International Chicken Genome Sequencing Consortium 2004b; Stiglec *et al.* 2007), we used genomic BLAST to ensure that the W contigs in our reference genome are representative of W-specific sequence. In addition, we confirmed any outliers in the initial W-specific set by comparing features of each W contig to features of the known set of autosomal and Z-linked contigs. We used 1000 bootstrap replicates to estimate confidence intervals of mean coverage and read depth for known autosomal or Z-linked contigs, which were then compared to the coverage and read depth values, respectively, of each putative W-specific contig. Our method is based on the assumption that very few ZZ reads should align to W-specific contigs, which as a result should have significantly lower coverage and read depth compared to autosomal or Z-linked contigs (Figure 3.1C). To confirm the predictions of our method, we compared the coverage and read depth for contigs of known location. We used nonpara-

metric bootstrapping methods to determine whether known W and known autosomal or Z-linked contigs had different distributions of coverage and read depth. For each of the 1000 bootstrap replicates, we calculated the difference between the 100th quantile of the W bootstrap distribution and the 0th quantile of the non-W-specific bootstrap distribution. This difference should be positive if the distribution of coverage or read depth of autosomal/Z-linked contigs is distinctly greater than that of W-specific contigs.

2.6.3 Simulations to determine effect of contig length

Because the length of unmapped contigs varied greatly (from 50 to 44,574 unmasked bp), we tested the effect of length by simulating genomes consisting of different-sized contigs. Contigs were sorted by length into 500 bp bins. We fragmented the mapped portion of the reference genome into contigs of length 500 bp, 1 kb, etc. For each fragmented genome, we redid the alignments and compared the distributions of coverage and read length for W- and non-W-specific contigs.

2.6.4 Classification approach

We developed a naïve Bayes classifier to identify W-specific contigs. A naïve Bayes classifier uses a set of training data to calculate the probability that a given example belongs to a certain class based on a set of features. It simplifies the learning process by assuming that the features are independent, although in practice it performs well even if that assumption is violated. We will refer

to each contig by its feature vector $X = (x_1, x_2)$, where x_1 is coverage and x_2 is read depth. The goal is to find the class C that maximizes the likelihood: $P(X|C) = P(x_1, x_2|C)$. C can be either W or non-W. Since we assume that x_1 and x_2 are conditionally independent, we can simplify this conditional probability to $P(X|C) = P(x_1|C)P(x_2|C)$. To account for the effect of contig length, we conditioned on length in the classification method as follows: given a contig X with length L (rounded to the nearest 500 bp), we rewrite the likelihood as $P(X|C, L) = P(x_1|C, L)P(x_2|C, L)$. The training set therefore depends on the contig length: for contigs of length L, the training set consists of mapped contigs of length L (see length simulations above). The feature probability distributions $P(x_i|C, L)$ are estimated from the relative frequencies of the appropriate training set. Both the coverage and read depth distributions were discretized into bins of equal width. We tested several bin widths: 0.0005, 0.001, 0.005, 0.01, and 0.05. Thus $P(a < x_i < b|C, L)$ is the frequency of contigs of class C with $a < x_i < b$ in the genome with length L contigs + ϵ, where ϵ is close to zero. This small sample correction is necessary because zero probabilities cause information loss. The posterior probability that a given contig is W-specific is then:

$$P(C \in W|X, L) = \frac{P(C \in W)P(X|C \in W, L)}{P(X|C \in W, L)P(X|C \in nonW, L)} \tag{2.1}$$

2.6.5 Performance

We assessed the performance of our test using Receiver Operating Characteristic (ROC) curves. ROC curves plot the true positive rate and false positive rate of a classifier over a range of threshold values, and the area under the curve (AUC) is a traditionally used statistic for model comparison. We generated ROC curves and calculated the AUC using the package ROCR in the R statisti-

cal package (http://www.r-project.org). A series of cross-validation tests using the previously-mapped contigs was used to fine-tune the bin sizes of classifier feature distributions and evaluate the effects of training set size and sample imbalance.

2.6.6 Validation and follow-up

W-specific candidates were verified using PCR. Genomic DNA was extracted from the blood of two female and two male White Leghorn chickens using the Qiagen DNeasy kit. Primers were designed for each candidate contig, and amplification was attempted in all four individuals (see Table A.1 for primer sequences and PCR conditions). If a given contig amplified successfully in both females but not in either male, then it was considered female-specific. Some candidates were verified via PCR in two female and two male Red Junglefowl (UCD 100 Red Jungle Fowl, from M.E. Delany, University of California, Davis). Primer pairs were scored for their ability to produce bands from both female templates that differed from the bands produced from both male templates. Primer pairs with identical results on male and female templates were scored as non-specific. The validation results were used in additional tests of performance. We used independent tests to further investigate the effects of contig length and sample imbalance on the predictive accuracy of our classifier. Validated W-specific candidates will be annotated in Bellott *et al.* in prep.

2.7 Acknowledgments

We thank Karel A. Schat for generously providing White Leghorn (N-2 line) tissue samples, as well as Alex Coventry and Wes Hochachka for helpful statistical discussions. Thanks to Clement Chow, Tim Connallon, Angela Early, and Scott Edwards for valuable comments on the manuscript. Grace Chi helped with labwork for some validations. NC was supported by a National Science Foundation Graduate Research Fellowship. This work was supported in part by NIH grant R01 GM64590 to AGC and AB Carvalho. This chapter is based on the manuscript: Chen N., Bellott D.W., Page D.C., Clark A.G. 2012. Identification of avian W-linked contigs by short-read sequencing. *BMC Genomics* **13**: 183.

CHAPTER 3

USING MENDELIAN INHERITANCE TO IMPROVE HIGH THROUGHPUT SNP DISCOVERY

3.1 Abstract

Restriction site-associated DNA sequencing or genotyping-by-sequencing (GBS) approaches allow for rapid and cost-effective discovery and genotyping of thousands of single nucleotide polymorphisms (SNPs) in multiple individuals. However, rigorous quality control practices are needed to avoid high levels of error and bias with these reduced representation methods. We developed the first formal statistical framework for filtering spurious loci using Mendelian inheritance patterns in nuclear families that accommodates variable-quality genotype calls and missing data - both rampant issues with GBS data - and for identifying sex-linked SNPs. Simulations predict excellent performance of both the Mendelian filter and the sex-linkage assignment under a variety of conditions. We further evaluate our method by applying it to real GBS data and validating a subset of high quality SNPs. These results demonstrate that our metric of Mendelian inheritance is a powerful quality filter for GBS loci that is complementary to standard coverage and Hardy-Weinberg filters. The described method, implemented in the software MendelChecker, will improve quality control during SNP discovery in non-model as well as model organisms.

3.2 Introduction

The advent of next-generation sequencing technologies has revolutionized biological research by allowing the pursuit of fundamental ecological and evolu-

tionary genomics questions in non-model organisms (Hudson 2008). It is now feasible to discover genome-wide markers in any species, even if little or no prior genetic resources are available (Ellegren and Sheldon 2008). However, many modern studies now require high quality genotypes for tens or hundreds of individuals. While recent technological advances have significantly lowered the cost of DNA sequencing, it is still expensive to assay genetic variation in large numbers of individuals (Narum *et al.* 2013).

Several methods have been developed to reduce the cost of high-throughput genotyping by restricting the complexity of the genome. A suite of these methods selectively sequence regions of the genome near restriction sites, allowing simultaneous discovery and genotyping of thousands of single nucleotide polymorphisms (SNPs) distributed across the genome. Several variations exist, but these methods are generally known as restriction site-associated DNA sequencing (RAD-seq) or genotyping-by-sequencing (GBS; reviewed in Davey *et al.* 2011). GBS methods have been used in a variety of applications, including phylogenetics (Rubin *et al.* 2012), population genomics (White *et al.* 2013), genome-wide association studies (Parchman *et al.* 2012), speciation genomics (Taylor *et al.* 2014), and genetic mapping (Andolfatto *et al.* 2011).

A central challenge in analyzing GBS data is the high variation in coverage across individuals and across loci, creating uncertainty in SNP calls and genotype assignments (Davey *et al.* 2011). In addition to the polymerase chain reaction (PCR) and sequencing error associated with next-generation sequencing platforms, this cost-effective method of high-throughput genotyping comes with its own set of caveats: restriction fragment length bias and PCR GC content bias contribute to high variation in read depth among loci, and restriction-site

polymorphism can skew allelic representation and therefore estimates of population genetic parameters (Arnold *et al.* 2013; Davey *et al.* 2013; Gautier *et al.* 2013). Spurious SNP calls may also result from collapsed paralogs or repeats during *de novo* assembly of reads into putative unique loci. Most GBS studies have used a set of heuristic criteria to filter out spurious sites, including read depth, proportion of missing data, and observed heterozygosity (Davey *et al.* 2011). While these simple filters are expected to discard most problematic loci during variant discovery, more sophisticated bioinformatic filtering tools are needed, especially since validation of large sets of SNPs remains prohibitively expensive (Narum *et al.* 2013). Here we present a novel framework for filtering spurious GBS loci based on a quantitative assessment of Mendelian errors in nuclear families.

Checking for Mendelian inheritance of genotypes has long been routine practice for removing genotyping errors in human linkage studies (Sobel *et al.* 2002), and there are multiple software packages that identify genotyping and pedigree errors (MENDEL, Stringham and Boehnke 1996; PedCheck, O'Connell and Weeks 1998; MERLIN, Abecasis *et al.* 2002; PLINK, Purcell *et al.* 2007). A recent study showed that imposing Mendelian inheritance constraints when assigning genotypes in parent-offspring trios results in higher genotyping accuracy and haplotype inference (Chen *et al.* 2013). To date, only a handful of GBS studies have used Mendelian inheritance as an additional filter. Most of these studies simply discarded any loci with extreme segregation distortion (Miller *et al.* 2012; Gagnaire *et al.* 2013; Ogden *et al.* 2013). Senn *et al.* (2013) used an estimate of Mendelian error rate to set a threshold for genotype confidence scores, but they only considered two cases of Mendelian error and did not incorporate genotype probabilities or sex-linkage into their estimates. Ignoring sex-linkage

is problematic because the different inheritance patterns of sex-linked sites may cause true sex-linked sites to be erroneously discarded as Mendelian errors under an autosomal model of inheritance.

Here we describe the first formal statistical framework that combines genotype probabilities with pedigree information to perform a quantitative analysis of Mendelian violation across sites and pedigrees and calculates the probability that a given SNP is sex-linked. Instead of focusing on individual genotyping errors, our goal is to evaluate the quality of putative variant sites during the SNP discovery process. Although we primarily discuss GBS data in this paper, our method, implemented in the C++ program MendelChecker, can be applied to any data set containing probabilistic genotype calls for at least one parent-offspring trio. The performance of MendelChecker on simulated and real data sets demonstrates that adding a Mendelian inheritance filter substantially improves the removal of spurious sites during SNP discovery.

3.3 Methods

3.3.1 Checking for Mendelian inheritance:

In diploid organisms, true nuclear genetic variants should follow patterns of Mendelian segregation in families, assuming no pedigree errors and no novel mutations in the offspring. In this paper we define Mendelian errors as genotypes that are inconsistent with their respective pedigree. Because genotyping errors may create Mendelian errors in otherwise legitimate segregating sites, it is important to consider genotype probabilities when evaluating Mendelian inheritance. We developed an efficient and scalable algorithm that iterates over all

possible genotypes in all individuals at a given site and calculates the likelihood of the pedigree given the genotype probabilities. Also, we use these pedigree likelihoods to evaluate the quality of each site and assign a probability of sex-linkage for each SNP.

For diploid individuals, there are 10 possible genotypes at each biallelic SNP. Based on the base calls from the sequence reads that overlap a given site, we assign each individual a vector of genotype probabilities: $(p_{AA}, p_{AC}, p_{AG}, p_{AT}, p_{CC}, p_{CG}, p_{CT}, p_{GG}, p_{GT}, p_{TT})$. Across the sample, only three genotypes should have non-zero probabilities for biallelic sites. By considering all 10 genotype probabilities, our method is flexible enough to accommodate multiallelic sites. We calculate the frequencies of all four alleles (p_A, p_C, p_G, p_T) from the observed genotype probabilities of all the parents (who are assumed to be unrelated) and impute a vector of expected genotype frequencies in the population:

$$G_{exp} = (p_A p_A, p_A p_C, p_A p_G, p_A p_T, p_C p_C, p_C p_G, p_C p_T, p_G p_G, p_G p_T, p_T p_T) \tag{3.1}$$

The statistical framework for our method was adapted from Jurg Ott's pedigree likelihood (Ott 1974). The classical pedigree likelihood consists of three components: $\text{Pen}(X_i|G_i)$, $\text{Prior}(G_j)$, and $\text{Trans}(G_o|G_f, G_m)$. $\text{Pen}(X_i|G_i)$ denotes the penetrance, or the conditional probability that individual i has an observed phenotype X_i given an unobserved genotype G_i. $\text{Prior}(G_j)$ is the probability that individual j has genotype G_j. Let $\text{Trans}(G_o|G_f, G_m)$ be the probability that two parents with genotypes G_f and G_m produce an offspring o with genotype G_o. The likelihood L of a pedigree with n individuals is:

$$L = \sum_{G_1} \cdots \sum_{G_n} \prod_i \text{Pen}(X_i|G_i) \prod_j \text{Prior}(G_j) \prod_{\{o,f,m\}} \text{Trans}(G_o|G_f, G_m) \tag{3.2}$$

where the product on j is taken over all parents, or founders, and the product on $\{o, f, m\}$ is taken over all parent-offspring trios.

In our case, the "phenotype" of interest is the true genotype; therefore, the penetrance function is equivalent to Prior(G_j). For a nuclear family with s offspring, the pedigree likelihood is reduced to:

$$L_A = \sum_{G_i} \prod_{i=1}^{s+2} \text{Prior}(G_i) \prod_{o=1}^{s} \text{Trans}(G_o | G_f, G_m) \tag{3.3}$$

Due to the high sampling variance of GBS data, not all individuals will be genotyped at all putative sites. If the genotype is missing for a particular individual, we substitute the expected genotype frequency (from G_{exp}) for Prior(G_i). In situations where the number of founders is too low to reasonably infer expected genotype frequencies in the population, we allow the option of using a uniform prior for missing genotypes.

To account for varying numbers of offspring in each nuclear family and variable minor allele frequency (MAF), we normalize the pedigree likelihoods for the number of informative trios in each family. We do so by dividing by the likelihood of a completely uninformative pedigree L_U, or the likelihood of the pedigree if the genotype probability vectors for all individuals were G_{exp} (the expected genotype frequencies). If there are insufficient unrelated individuals sampled to accurately estimate population allele frequencies, we can calculate L_U using a uniform prior. We combine individual pedigree likelihoods for all n pedigrees to obtain a single composite score for each site, M:

$$M = \sum_{i=1}^{n} \log \frac{L_{Ai}}{L_{Ui}} \tag{3.4}$$

Note that our metric for quantifying the degree of Mendelian inheritance, M, incorporates several factors. The highest scoring sites will have high quality

genotype calls in multiple individuals, a low rate of missing data, and a large proportion of genotype configurations consistent with Mendelian transmission.

3.3.2 Assessing the probability of sex-linkage:

Sex-linked sites have different transmission probabilities compared to autosomal sites (Elston and Stewart 1971). Some true sex-linked sites would erroneously appear as Mendelian errors under an autosomal model of inheritance. Thus, for each SNP, we calculate pedigree likelihoods and M under both an autosomal model of inheritance and a sex-linked model of inheritance. Transmission probabilities for sex-linked sites depend on the sex of the offspring. Therefore, the likelihood of the pedigree under a sex-linked model incorporates the sex of each offspring o:

$$L_S = \sum_{G_i} \prod_{i=1}^{s+2} \text{Prior}(G_i) \prod_{o=1}^{s} \text{Trans}(G_o | G_f, G_m, sex_o) \tag{3.5}$$

If the sex of the offspring is unknown, we take the average of the male and female transmission probabilities. We use the combined pedigree likelihoods to compute the posterior probability that a given site is sex-linked, S, using Bayes' Theorem:

$$S = \frac{\alpha \sum_{i=1}^{n} \log L_{Si}}{\alpha \sum_{i=1}^{n} \log L_{Si} + (1-\alpha) \sum_{i=1}^{n} \log L_{Ai}} \tag{3.6}$$

where α is the prior probability of sex-linkage and is estimated as the proportion of the genome on the X or Z chromosome, and n is the number of pedigrees. To evaluate each SNP, we first classify the site as autosomal or sex-linked based on S and then evaluate the SNP with the appropriate M.

3.3.3 Simulations to assess performance:

We performed a series of simulations to assess the performance of our method under different scenarios. We used custom Perl scripts to generate genotype probability vectors for nuclear families of varying offspring number, Mendelian error rate, proportion of missing data, MAF, and genotype quality for both sex-linked and autosomal sites. Normalized, Phred-scaled genotype likelihood scores (similar to the PL field in VCF files) were simulated based on an exponential distribution with means estimated from real data (see below). We used a mean of 3000, 500, and 100 for high-, medium-, and low-quality genotypes, respectively. For a high quality site, the most likely genotype was assigned a Phred-scaled likelihood of 0, the second most likely genotype was sampled from an exponential distribution with mean 300, and the third most likely genotype was sampled from an exponential distribution with mean 3000. Mendelian errors were introduced by forcing an offspring to have a genotype that would be inconsistent with Mendelian transmission given the parental genotypes. We assumed a 50:50 sex ratio when assigning sex to each offspring. Unless otherwise specified, we simulated 5,000 autosomal and 5,000 sex-linked SNPs with MAF of 0.05 or 0.25 in Hardy-Weinberg genotype proportions for each scenario and ran our simulated data through MendelChecker.

To verify the functionality of our Mendelian SNP score, we simulated 10 parent-offspring trios and varied the proportion of the families containing a Mendelian error from 0 to 1. All SNPs had high genotype quality and no missing data, allowing us to assess the sensitivity of MendelChecker under ideal conditions. We first evaluated the extent to which we could assign sex-linkage from pedigree likelihoods by comparing individual autosomal and sex-linked

pedigree likelihoods. To determine the robustness of our sex-linkage posterior probability to the prior, we ran MendelChecker on the same dataset using different prior probabilities of sex-linkage (0.1, 0.5, and 0.9).

Above, we assessed the performance of MendelChecker given a known number of Mendelian errors, assuming that all sampled trios were informative. However, in real data, not all genotype errors will lead to inconsistent pedigrees. To more realistically assess our power to detect spurious SNPs, we simulated genotyping errors that are consistent with Mendelian transmission. In these simulations, we introduce spurious sites by simulating offspring genotypes independent of the parental genotypes, *i.e.*, as if they were unrelated. The next set of simulations focused on testing the power to detect Mendelian errors under different genotype qualities, proportion of missing data, and MAF in 10 parent-offspring trios. The full range was tested for each parameter: we varied genotype quality from low (mean phred score of 100) to high (mean phred score of 3000), the fraction of missing data from 0 to 1, and the MAF from 0.01 to 0.5.

We examined the influence of sampling scheme by simulating nuclear families of different sizes. First, we assessed performance as we sample increasing numbers of parent-offspring trios. Then, we held the number of meioses constant and changed the family configuration: we compared results for 10 trios, 5 families with 2 offspring each, 2 families with 5 offspring each, and 1 family with 10 offspring. We assessed the power of MendelChecker when samples include both parents, only the homogametic parent, or only the heterogametic parent. For these simulations, we generated SNPs with medium to high quality genotypes and 0-20% missing data.

We estimated the ability of MendelChecker to assign sex-linkage or to de-

tect Mendelian errors in each scenario by generating Receiver Operating Characteristic (ROC) curves and calculating the area under the curve (AUC) using the package ROCR (Sing *et al.* 2005) in the R statistical package (http://www.r-project.org).

3.3.4 Data collection:

We validated our method using data obtained from a long-studied population of Florida Scrub-Jays (*Aphelocoma coerulescens*) from Archbold Biological Station. Florida Scrub-Jays are genetically monogamous (Townsend *et al.* 2011); therefore we can confidently assume that the pedigrees constructed from field observations are accurate. We sampled 103 individuals in 27 nuclear families from 1989 and 2008. Genomic DNA was extracted from blood samples stored in lysis buffer using the Qiagen DNeasy kit. We slightly modified the GBS protocol of Elshire *et al.* (2011) to generate multiplexed reduced representation libraries for Illumina sequencing. Briefly, 500 ng of DNA from each individual was digested with the enzyme *Pas*I (NEB) before ligation of barcoded adapters. Individual samples were then pooled and cleaned using a Qiagen Minelute PCR purification kit. Libraries were amplified with PCR with short extension times to favor amplification of shorter fragments (98°C for 30 s; 18 cycles of 98 °C for 30 s, 65°C for 7 s, 72°C for 7 s; 72°C for 5 min). Final library cleanup was performed with AMPure XP beads (Agencourt). We generated multiplexed libraries consisting of 6 to 12 individuals per lane and sequenced 5 lanes on the Illumina GA II (84 and 86 bp reads) and 8 lanes on the Illumina HiSeq 2000 (56 and 101 bp reads). Three libraries were sequenced twice. Sequencing was done at the Cornell University Biotechnology Resource Center Genomics Facility and the Weill Cornell

Genomics Resources Core Facility. All sequence data have been submitted to the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession SRP041511.

3.3.5 Pipeline for obtaining probabilistic genotype calls:

Because it is crucial for our downstream analysis to obtain probabilistic genotype calls and propagate error throughout the analysis, we created a flexible analysis pipeline for calling genotypes from GBS data (Figure 3.1). We used custom Perl scripts to sort the raw reads by barcode into individual files and trim off adapters and low quality bases. These demultiplexed and processed reads were all trimmed to 79 bp.

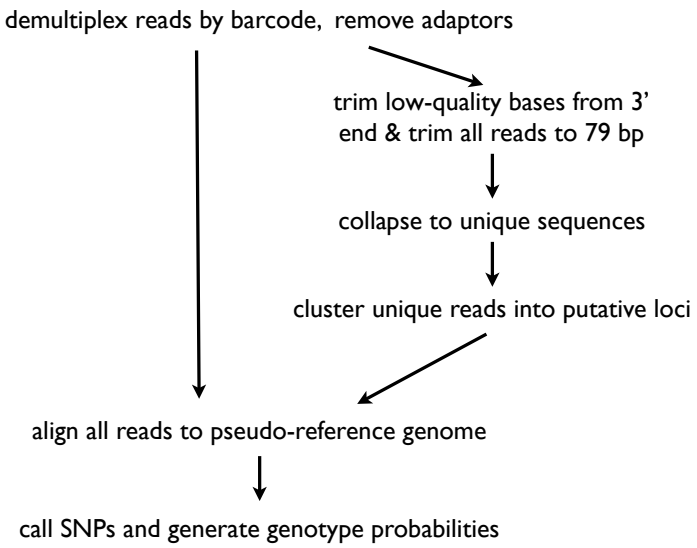

Figure 3.1: Overview of our custom pipeline for obtaining probabilistic genotype calls from GBS or RAD-seq data.

To take advantage of well-established software for variant detection and genotyping in the absence of a reference genome, we generated a "pseudo-reference genome" that contains all sites in the genome sampled in this reduced representation sequencing approach. We took the entire set of sequences from all individuals and collapsed the reads into a set of unique sequences, removing singletons in the process. We used the program SlideSort (Shimizu and Tsuda 2011) to perform a rapid all-by-all pairwise comparison of all unique reads and generate a list of all pairs that differ by 10 bp or less. Reads were grouped into clusters with the Markov Cluster Algorithm (MCL; van Dongen 2000), which allows the formation of clusters with multiple SNPs. A single sequence from each cluster was included in the pseudo-reference genome.

We retained and used base quality scores from the original reads. We aligned the processed reads from each individual to the pseudo-reference using BWA (Li and Durbin 2009). BAM files were sorted and merged with Picard tools (http://picard.sourceforge.net) before indel realignment and variant calling with the UnifiedGenotyper in GATK (DePristo *et al.* 2011). GATK performs SNP discovery and probabilistic genotype calling across all samples simultaneously, which is advantageous because multiple-sample variant calling is more accurate than calling SNPs in each individual separately (Nielsen *et al.* 2011).

3.3.6 Validation on real data:

We ran MendelChecker on the resulting VCF file to assess Mendelian inheritance. In the 1.2 Gb Zebra Finch genome, the Z chromosome is ~73 Mb in length (Warren *et al.* 2010). Assuming the Florida Scrub-Jay has similar chro-

mosome sizes as the Zebra Finch, we used a prior probability of sex-linkage of 0.06. Using VCFtools (Danecek *et al.* 2011), we removed individual low-quality genotype calls (GQ<20) before calculating statistics about each site. We applied a series of stringent filters, removing all sites with low mapping quality or read depth (MQ<35, QD<5) or high levels of missing data (>20%). From this set of higher-quality SNPs, we either (1) removed sites that deviated from Hardy-Weinberg proportions in the 50 founders (p<0.001) or had a high proportion of heterozygote calls (>75%), (2) removed sites with M<-10, or (3) applied both filters. We selected 1,160 high quality SNPs for genotyping in 96 individuals using custom Illumina iSelect Beadchips. The genotyping accuracy of iSelect BeadChips exceeds 99% (Steemers and Gunderson 2007), and here we use these BeadChip genotypes as our validation set. We calculated genotype concordance as the proportion of maximum likelihood genotypes from the GBS data that match the BeadChip genotype calls. All genotype data have been submitted to dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) under accession numbers NCBI ss995818232-995820422.

To test the ability of our method to predict sex-linkage, we used a two-step approach to determine the putative chromosomal location of high quality SNPs. First, we aligned the pseudo-reference genome (the collection of all sampled loci) to the Florida Scrub-Jay draft genome (Chen *et al.* in prep) using BWA. The Florida Scrub-Jay genomic scaffolds were aligned to the Zebra Finch genome using standalone BLAST (Camacho *et al.* 2009) and assigned to putative chromosomes based on the best BLAST hit. The robustness of this annotation method relies on the high degree of synteny among extant bird lineages (Ellegren 2010). We calculated the AUC value using these chromosomal assignments. All analyses were done in the R statistical package.

3.3.7 Implementation:

Our method for checking for Mendelian inheritance, MendelChecker, has been implemented in C++ and is available for download at http://sourceforge.net/projects/mendelchecker/. Scripts and instructions for our custom pipeline for obtaining probabilistic genotype calls from GBS data are available upon request.

3.4 Results

3.4.1 Simulations:

We used simulations to (1) verify the accuracy of our quantitative framework for assessing Mendelian violations and (2) evaluate the performance of our metrics (S and M) under different scenarios. In the initial verification step, we simulated SNPs with no missing data and high quality genotype calls in 10 trios with varying amounts of Mendelian error. As the proportion of families containing a Mendelian error increases, M decreases. M is lower for SNPs with lower MAF (Figure 3.2A). For autosomal SNPs, the pedigree likelihood calculated under an autosomal model of inheritance is greater than the likelihood calculated under a sex-linked model of inheritance, and vice-versa for sex-linked SNPs (Figure B.1). Therefore, we can use pedigree likelihoods to calculate the posterior probability that a given SNP is sex-linked. The posterior probability of sex-linkage is an accurate classifier. For SNPs with MAF 0.25, AUC values range from 0.99 to 0.91 for SNPs with no Mendelian errors and 5 Mendelian errors, respectively. As the number of Mendelian errors increases, our ability to distinguish sex-linked SNPs from autosomal SNPs decreases (Figure 3.2B). We calculated S for the same dataset using different prior probabilities of sex-linkage and found

that the AUC values changed by less than 0.002 in all cases. Thus, our model is robust to the prior for these simulation parameters.

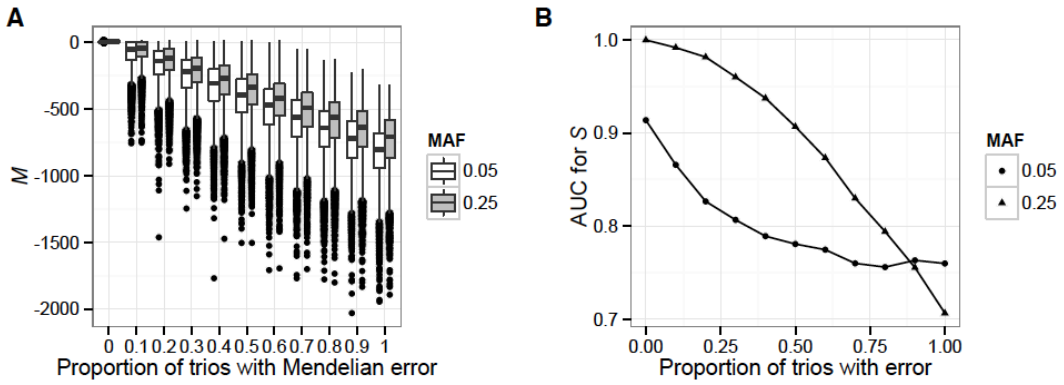

Figure 3.2: Verifying the assumptions underlying MendelChecker. We simulated 5,000 autosomal and 5,000 sex-linked SNPs for 10 offspring trios and varied the proportion of families containing a Mendelian error, *i.e.*, an error that is inconsistent with the pedigree. Here we show results for sites with MAF of 0.05 and 0.25. **(A)** Boxplots for M, our metric of Mendelian inheritance. Results for SNPs with MAF 0.05 are shown in white, and SNPs with MAF 0.25 are shown in gray. M decreases as the proportion of families containing a Mendelian error increases. **(B)** Performance of the sex-linkage classifier. Here, points indicate the MAF of the SNPs. The AUC value for S decreases as the proportion of trios with a Mendelian error increases.

This first set of simulations used ideal conditions - high genotype quality and no missing data. However, these conditions are rarely met in real data. Because the pedigree likelihoods are influenced by genotype quality and proportion of missing data as well as MAF, our next set of simulations explored the relative contribution of these other factors to S and M. We generated spurious SNPs by simulating offspring genotypes independently of the parental genotypes, allowing both Mendelian consistent and inconsistent errors. We simulated true and spurious SNPs in 10 trios, and systematically varied genotype quality, the proportion of missing data, and MAF. MendelChecker can correctly assign sex-linkage under almost all conditions. AUC values for S decrease below 0.9 only when the proportion of missing data exceeds 0.7 or the MAF is 0.01 (Figure 3.3). As expected, as genotype quality decreases and the proportion of missing data increases, M increases for spurious SNPs, indicating lower probability of detecting Mendelian errors (Figure 3.3A-F). AUC values for M stay above 0.98 for medium genotype qualities but drop to 0.81 for low quality genotypes (Figure 3.3C). Missing data has a larger impact on the performance of MendelChecker; our ability to detect a Mendelian error decreases with the proportion of missing data, with AUC below 0.9 when more than 20% of the individuals have missing genotypes (Figure 3.3F). The ability to detect errors is low for SNPs with MAF<0.05 because most individuals are homozygous for the major allele, resulting in few Mendelian inconsistent errors (Figure 3.3G-I).

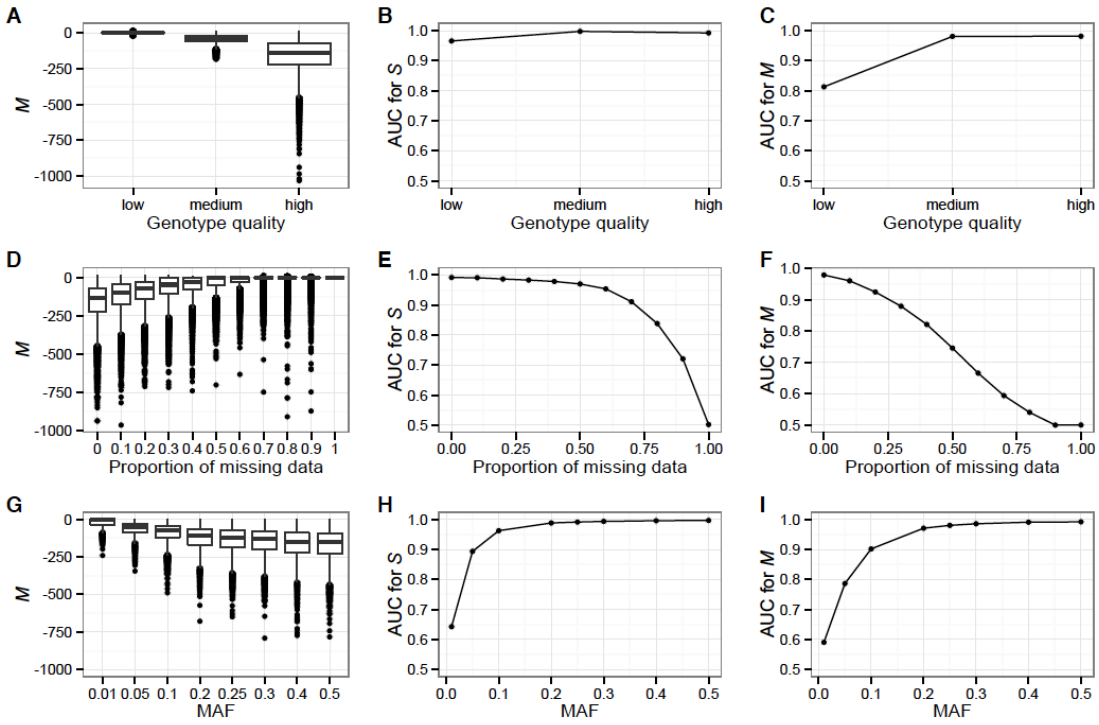

Figure 3.3: The influence of confounding factors on the ability to identify errors. Here we used a sample of 10 parent-offspring trios and simulated 10,000 SNPs with no errors and 10,000 SNPs with error. The left column shows M values for spurious SNPs, the middle column plots AUC values for S, and the right column shows AUC values for M. Note that higher M values for spurious sites indicates a decreased power to detect error. **(A-C)** Genotype quality has a minimal effect on the performance of MendelChecker. **(D-F)** As the proportion of missing data increases, the ability to assign sex-linkage and detect spurious sites decreases. **(G-I)** MendelChecker has decreased performance for SNPs with very low MAF (<0.1). The AUC for M drops below 0.9 only for low quality genotypes, greater than 20% missing data, or MAF<0.1.

After characterizing the influence of Mendelian errors as well as data quality and missingness on both S and M, we assessed the performance of MendelChecker for different sampling schemes. The power to assign sex-linkage and detect error increases as we sample more trios (Figure 3.4AB). Sex-linkage assignment is accurate (AUC>0.9) with a sample size of 4 trios when MAF=0.25 (Figure 3.4A). The AUC for M exceeds 0.90 with just 10 trios for SNPs with MAF=0.25 (Figure 3.4B). More trios (25) are needed to achieve an AUC>0.90 for rare SNPs (MAF=0.05; Figure 3.4AB). Given a set number of trios, we tested whether it is better to sample more families with fewer offspring each or fewer families with more offspring each. We simulated several possible family configurations when sampling 10 trios and found that it is more advantageous to sample multiple smaller families (Figure 3.4CD). In all cases, missing one parent decreases the AUC for M, though, as expected, performance of the sex-linkage assignment is lower only if the heterogametic parent is missing.

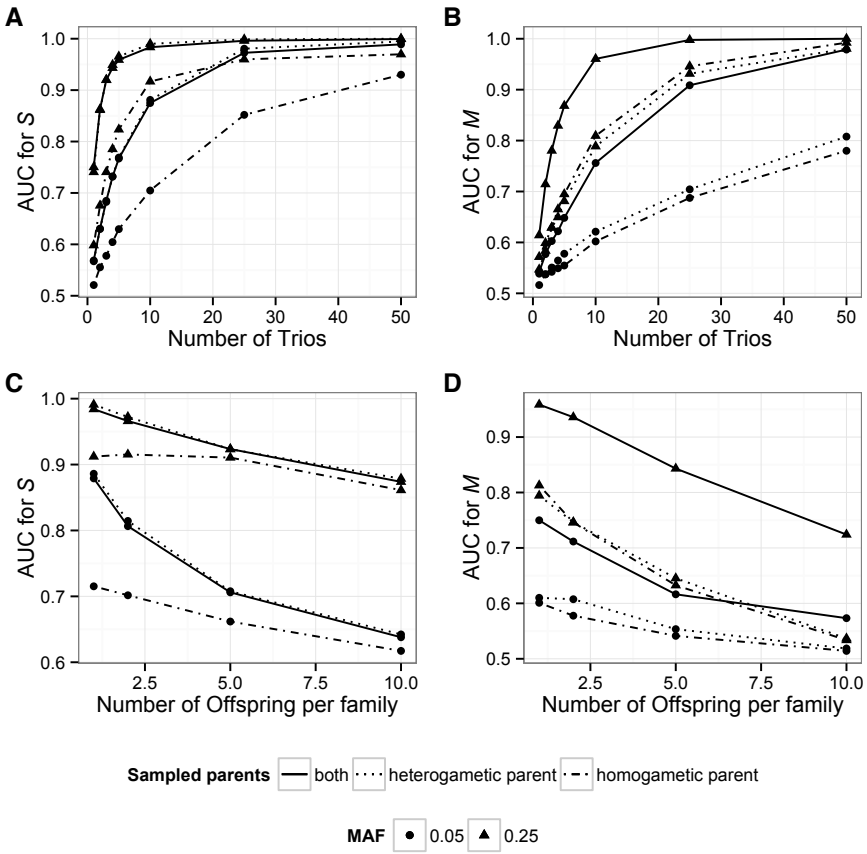

Figure 3.4: The influence of sampling scheme on the ability to identify spurious variant sites based on Mendelian errors. Here, lines indicate whether both parents or only one parent is sampled, and the points indicate the MAF of the SNPs. **(A-B)** The power to assign sex-linkage and identify spurious SNPs increases as the number of sampled parent-offspring trios increases. **(C-D)** The configuration of the families also affects power. Given a set number of trios, AUC values are higher for a sample of 10 families with 1 offspring each compared to a sample of a single family with 10 offspring. Sampling more founders increases the probability of sampling informative trios. As expected, performance of MendelChecker is lower when only one parent is sampled.

3.4.2 Real data analyses:

We tested the performance of our method on GBS data collected from 103 Florida Scrub-Jays in 27 nuclear families. Illumina sequencing produced a total of 935,765,768 reads, of which 814,341,664 contained a unique barcode and minimal adapter sequence contamination. Our custom pipeline identified 266,806 biallelic SNPs. Distributions of various quality metrics for the full SNP set can be found in Figure B.2. However, after filtering on individual genotype quality and overall per-site quality, only 20,347 SNPs were genotyped in >80% of our individuals. Of these SNPs, 11,758 passed our Mendelian inheritance filter (M>-10), 19,241 passed a HWE test (p>0.001), and 10,855 passed both filters. In this case, M is a more conservative filter: 43.6% of the SNPs that pass the HWE test fail MendelChecker at this threshold for M. Applying a HWE filter after filtering based on M eliminates 7.7% of the Mendelian SNPs, all of which have MAF>0.07 (Figure B.3). MendelChecker is a more powerful filter than HWE for rare variants, but HWE performs better for sites with high MAF (Figure 3.5). At high MAF, the probability of two heterozygous parents increases, which in turn decreases the probability an error can be detected as a Mendelian inconsistency. For a biallelic site, two heterozygous parents can produce offspring with all four genotype configurations; therefore only errors that introduce a novel allele would be inconsistent with Mendelian inheritance patterns. It is important to consider different models of inheritance: 62.2% of putative sex-linked SNPs would have failed the Mendelian inheritance test under an autosomal model of transmission.

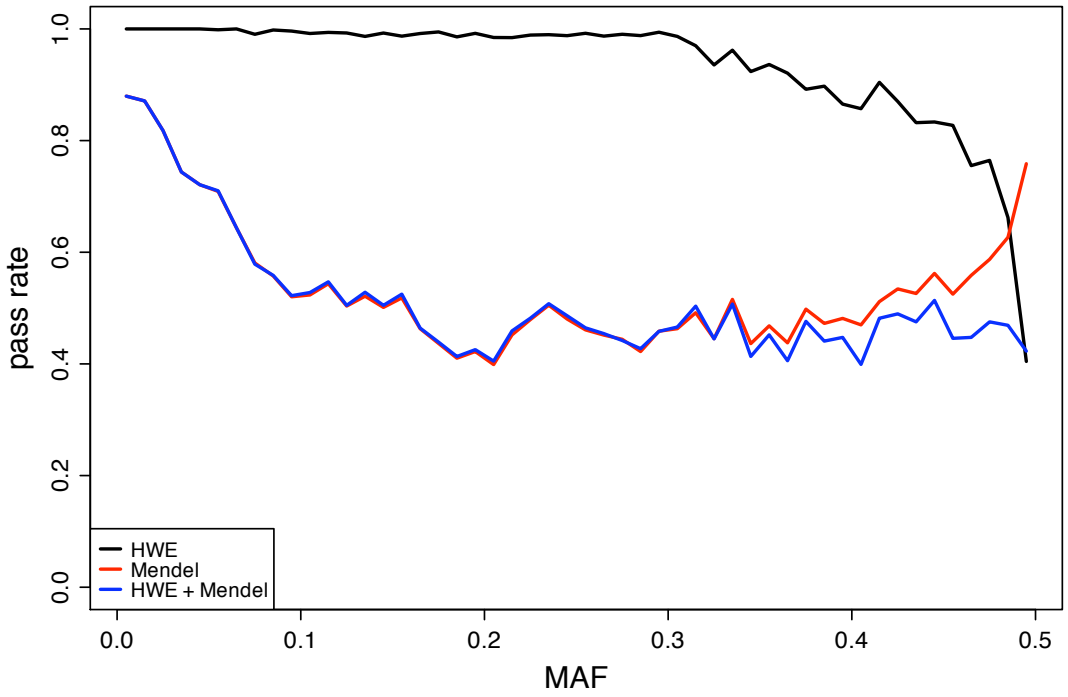

Figure 3.5: The proportion of SNPs with different MAFs that pass different QC filters. The HWE test (black) has low power to reject the null for SNPs with MAF<0.3. Mendelian inheritance (red) can filter out low frequency variants but loses power for variants with high MAF. At high MAF, the probability that both parents are heterozygous increases, and fewer errors can be detected as Mendelian inconsistencies. A combination of HWE and Mendelian inheritance tests (blue) can filter erroneous SNPs of all MAF. Comparison of pass rates as a function of MAF shows the complementary nature of the two filters.

We validated the genotype calls for 96 individuals at 1,160 SNPs using custom Illumina iSelect Beadchips. Mean genotype concordance is high (98.2%), and only 5.9% of these SNPs have genotype concordances lower than 95%. These SNPs are all high quality (QD>5, MQ>35, <20% missing data), consistent with HWE, and have relatively high M scores (M>-10). If we consider only the 686 SNPs with M>0, mean concordance increases to 98.7%, and the percentage of SNPs with concordance values below 95% drops to 3.2%. We acknowledge that an ideal validation experiment would have included low-quality SNPs. However, the high concordance of our validation set indicates that the coverage, HWE, and Mendelian inheritance filters we applied were successful in eliminating spurious sites.

Using alignment to the Zebra Finch genome, we were able to reliably assign putative chromosome locations to 7,744 of the 10,855 SNPs that passed all filters, with a minimum of 33 SNPs on every chromosome except Chromosome 16. The posterior probability of sex-linkage proved to be a reliable classifier: <0.2% of autosomal SNPs and 59% of Z-linked SNPs were classified as sex-linked (Figure 3.6) with an AUC of 0.85. Note that not all genotype configurations have different sex-linked and autosomal patterns of transmission, so it is not possible to identify all sex-linked SNPs based on a finite number of pedigrees. We suspect that the nine autosomal SNPs with high probabilities of sex-linkage could have been aligned to the wrong chromosome.

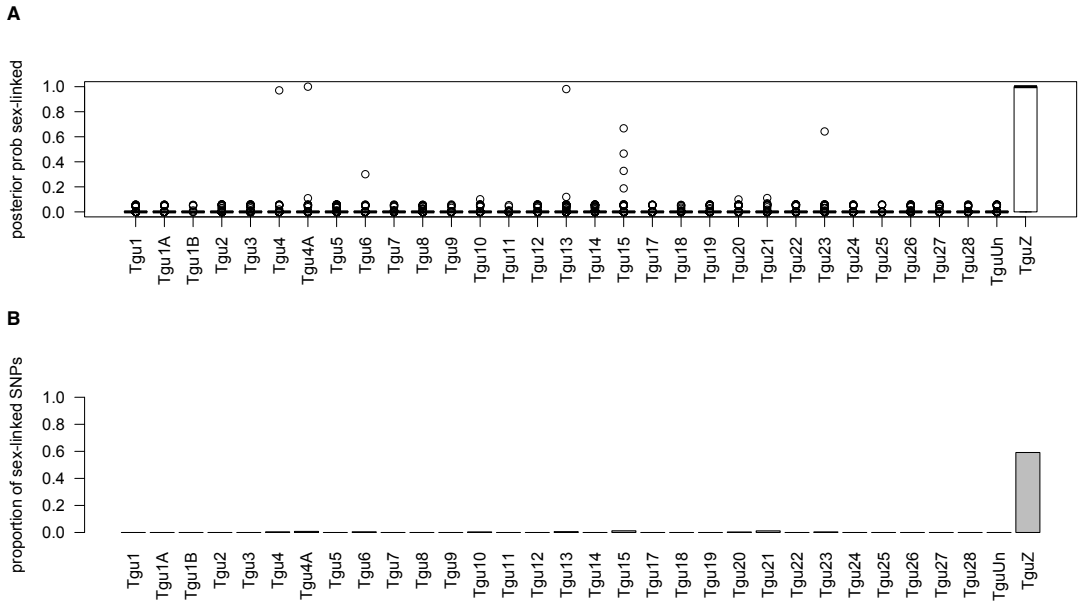

Figure 3.6: Assessment of our ability to assign sex-linkage. (A) Box-plots of the posterior probability of sex-linkage for SNPs on each chromosome. **(B)** The proportion of SNPs on each chromosome classified as sex-linked. Because not all genotype configurations have different autosomal and sex-linked transmission probabilities in our sample set, we do not expect to be able to classify all SNPs on the Z chromosome as sex-linked.

3.5 Discussion

GBS has become a popular approach for a myriad of ecological and evolutionary studies, but more advanced methods are required for quality control of SNP discovery using GBS, especially since GBS genotype calls are rarely validated. Here we present a novel framework for filtering spurious GBS loci based on a quantitative assessment of Mendelian errors in nuclear families and evaluate the performance of our method using simulated and real data. This is one of the few GBS studies to date to validate genotype calls using a different genotyping platform. MendelChecker assigns each site a probability of being sex-linked, S, and a quantitative score of Mendelian consistency, M. Users can use S to classify each site as putatively autosomal or sex-linked before ranking SNPs with the appropriate M score and specifying a threshold to identify spurious sites.

To obtain the highest quality set of SNP calls, we recommend combining M with other widely-used quality control measures, such as coverage and Hardy-Weinberg filters. Our simulations show that our power to detect a single Mendelian error decreases as the information content of the genotype data decreases. The ability to detect spurious sites is relatively poor (AUC<0.9) when all genotypes have low quality or when the proportion of missing genotypes is greater than 0.2. Many previous GBS studies routinely filter out sites with low quality or >20% missing data. Applying these standard genotype quality and coverage filters will remove sites with low information content. In analyzing GBS data collected from 103 Florida Scrub-Jays, we show that MendelChecker and HWE are complementary tests (Figure 3.5). A significant advantage of our method is that it can detect errors in rare variants, whereas HWE has low power to reject SNPs with low MAF. The performance of Mendelchecker is lower at

high MAF, in part because the probability of detecting a genotyping error as a Mendelian inconsistency is greater when the MAF is below 0.5 (Douglas *et al.* 2002). The fact that Mendelian inheritance patterns can provide no information about the validity of the SNP if both parents are heterozygous can be problematic when trying to identify spurious SNPs generated by collapsed paralogs: if every individual is heterozygous at a site, MendelChecker will assign it a high M score. Therefore, we recommend filtering based on both Mendelian inheritance and Hardy-Weinberg to remove spurious sites at all MAF.

The genomic locations of GBS loci are unknown in organisms without a closely related reference genome, and sex-linked loci are often of special interest. For instance, rates of evolution differ between sex chromosomes and autosomes, and sex-linked genes are thought to play an important role in speciation (Charlesworth *et al.* 1987; Presgraves 2008; Qvarnström and Bailey 2009). Thus the ability to classify SNPs as putatively autosomal or sex-linked expands the scope of questions that can be answered with GBS data. In addition to assessing Mendelian violations, MendelChecker calculates the posterior probability that a site is sex-linked. Our software can accommodate XY, ZW, and XO sex determination systems and can accurately assign sex-linkage to simulated and real SNPs. The transmission probabilities of some sex-linked SNPs differ from those of autosomal SNPs; therefore, assuming an autosomal pattern of inheritance for all loci may lead one to discard perfectly valid sex-linked SNPs. However, not all genotype configurations have different sex-linked and autosomal patterns of transmission, so MendelChecker does not have the ability to identify all sex-linked SNPs given finite numbers of pedigrees. For example, in organisms with pseudoautosomal regions, SNPs in those regions cannot be distinguished from autosomal SNPs. Future versions of MendelChecker could incorporate tests for

other markers with unusual inheritance patterns, such as mitochondrial DNA or chloroplast DNA.

Despite additional difficulty in obtaining pedigree data, sampling pedigrees has many advantages beyond improving SNP discovery: pedigrees are key to answering several fundamental questions in evolutionary biology (Kruuk and Hill 2008; Pemberton 2008). Pedigree information can be obtained by performing crosses, observing mating or parental care in captive or wild populations, or by collecting gravid females (Pemberton 2008). Several software programs to assign parentage based on highly variable genetic markers have been developed (reviewed in Blouin 2003; Jones and Ardren 2003). Of course, not all GBS experiments will consist solely of family groups. For experiments that require sampling multiple unrelated individuals, the inclusion of just four parent-offspring trios is sufficient to allow some filtering based on Mendelian inheritance for SNPs with MAF>0.05 (AUC>0.80). In this case, the multiple unrelated individuals can be used to estimate population allele frequencies, and SNPs can be filtered based on inheritance patterns in the included nuclear families. For a given sample size, there is a trade-off between sampling families and sampling additional unrelated individuals, but the advantage of obtaining a more accurate set of variant calls may be worth the slightly decreased sample size.

Currently, MendelChecker only considers nuclear families. Extended pedigrees can be broken into several separate nuclear families. While linkage map construction benefits greatly from multigenerational families, nuclear families are sufficient for identifying spurious SNPs based on Mendelian violations. Power to identify spurious SNPs based on Mendelian inheritance increases as more parent-offspring trios (meioses) are sampled. This is consistent with previ-

ous work showing that genotyping additional siblings in a family increases the genotyping error detection rate by 10-13%, depending on the allele frequencies of the variant (Gordon *et al.* 2000). Given a set number of trios, our simulations showed improved performance when multiple smaller families were sampled instead of fewer large families. Including more pairs of parents increases the probability of sampling informative parental genotype combinations.

Although the MendelChecker framework assumes accurate pedigrees, one can sum the pedigree likelihoods over all or a subset of high-confidence SNPs in order to identify pedigrees with disproportionally high rates of Mendelian error or to test alternative pedigrees. This alternative use of MendelChecker can prove especially useful in organisms for which parental assignments are uncertain, *e.g.*, birds with extra-pair paternity. However, the primary goal of MendelChecker is to identify high quality sites. Given a set of high-quality genotypes, other software packages exist for identifying potential pedigree errors (*e.g.*, PedCheck, O'Connell and Weeks 1998; RELPAIR, Epstein *et al.* 2000; PLINK, Purcell *et al.* 2007).

There are a number of well-developed applications for *de novo* analysis of GBS data, such as Stacks (Catchen *et al.* 2011), Peterson's ddRAD pipeline (Peterson *et al.* 2012), UNEAK (Lu *et al.* 2013), RApiD (Willing *et al.* 2011), pyRAD (Eaton 2014), RADtools (Baxter *et al.* 2011), and Rainbow (Chong *et al.* 2012). We developed a custom pipeline for additional flexibility and full control over the parameters at each step of the process. Compared to UNEAK, the most widely-used reference-free pipeline designed specifically for the Elshire *et al.* (2011) GBS method, our pipeline is less conservative when creating clusters and therefore more appropriate for systems with higher nucleotide diversity. MendelChecker

is compatible with any pipeline that provides posterior genotype probabilities.

In conclusion, we have designed a flexible quantitative test for Mendelian inheritance that propagates genotype uncertainty, accommodates missing data, is powerful for rare variants, and distinguishes between autosomal and sex-linked SNPs. We recognize that including families in population-scale datasets may require additional effort; however, we argue that future studies would benefit from including a subsample of nuclear pedigrees when possible because filtering based on Mendelian inheritance will result in a more accurate set of variant sites. Performance of MendelChecker increases as more meioses and more families are sampled, but the inclusion of 10 trios is sufficient for high performance (AUC>0.90). MendelChecker provides a statistical test for Mendelian errors and identifies sex-linked loci, making it a valuable tool for researchers using GBS data to explore ecological and evolutionary questions.

3.6 Acknowledgments

We thank Rob Elshire, Jen Grenier, Charlotte Acharya, Qi Sun, and Harpreet Singh for help troubleshooting GBS labwork and bioinformatics. The Florida Scrub-Jay samples and pedigrees were obtained thanks to John Fitzpatrick, Reed Bowman, Raoul Boughton, Shane Pruett, Laura Stenzler, and many students, interns, and staff at Archbold Biological Station. Thanks to the Clark and Harrison labs for comments. This work was supported by NSF (SGER DEB 0855879 and DEB 1257628), a Cornell Center for Vertebrate Genomics Seed Grant, the Andrew W. Mellon Student Research Award, and the Cornell Lab of Ornithology Athena Fund. N.C. was supported by a NSF Graduate Research Fellowship and a Cornell Center for Comparative and Population Genomics

Fellowship.

CHAPTER 4

REGIONAL POPULATION DECLINE IS ASSOCIATED WITH INBREEDING AND HATCH FAILURE IN THE FLORIDA SCRUB-JAY

4.1 Abstract

The population genetic consequences of declining population size are well described theoretically, but thorough empirical studies that disentangle the relative impacts of different evolutionary forces at a genome-wide scale in the wild are scarce because they demand huge field and laboratory investments. Analysis of time-series data instead of a single sampling time point can provide a more complete picture of the evolutionary processes influencing levels of genetic variation in a population. We characterized changes in genetic diversity over 23 years in the federally threatened Florida Scrub-Jay (*Aphelocoma coerulescens*), which has declined in number by at least 97% during the past 100 years, and by nearly 50% over the past 20 years. A population of Florida Scrub-Jays at Archbold Biological Station has been studied intensively since 1969, resulting in detailed phenotypic and demographic data from thousands of pedigreed individuals. We used custom Illumina iSelect Beadchips to genotype every nestling and immigrant recruited in our study population from 1988-1995 and 1999-2013 at 7,404 autosomal SNPs in approximate linkage equilibrium. Although our study population has remained stable in size through time, the decreasing proportion of breeders that are immigrants over time is correlated with an increasing mean inbreeding coefficient of the birth cohort. We find evidence for inbreeding among resident breeding pairs but not resident-immigrant pairs. Increasing relatedness among breeders has negative fitness consequences: observed patterns of hatching failure are best explained by levels of Identity-by-

Descent (IBD) sharing between parents. This study is one of the most detailed longitudinal investigations of genetics in a wild population to date, and the fine-scale calibration of impacts of declining population size will have significant management consequences.

4.2 Introduction

Unprecedented numbers of species are undergoing severe population declines worldwide, yet many details of the eco-evolutionary responses to population decline remain poorly characterized in wild species (Kohn *et al.* 2006). A major cause of population declines worldwide is human-mediated habitat fragmentation (Henle *et al.* 2004). Increased isolation of natural populations can lead to lower gene flow among populations and consequently decreased genetic diversity (Young *et al.* 1996). The prevalence of habitat fragmentation means that many natural populations exist as metapopulations, and investigations of the genetic consequences of population size fluctuations need to account for gene flow (Keller *et al.* 2001).

A number of studies have demonstrated the importance of gene flow in introducing genetic variation to small populations or governing genetic differentiation among subpopulations. Empirical support exists for rapid and substantial losses in fitness after the cessation of gene flow (Hogg *et al.* 2006). In small populations, a single immigrant can restore genetic variation originally lost to drift and reduce levels of inbreeding (Ingvarsson 2001). This phenomenon has been termed "genetic rescue" (Ingvarsson 2001). Even modest rates of immigration can lead to significant outbreeding and population growth (Vilà *et al.* 2003) as well as rapid restoration of genetic diversity after a population bottle-

neck (Keller *et al.* 2001). Genetic rescue has been documented in several populations (classic examples include Prairie Chickens, Bighorn Sheep, and Florida panthers), but the beneficial effects may be short-lived and/or masked by environmental conditions (Adams *et al.* 2011). Population viability analyses and conservation management plans should aim for long-term genetic restoration (Adams *et al.* 2011). However, longitudinal studies of genetic diversity in a single population through time remain rare (Vilà *et al.* 2003; Kaeuffer *et al.* 2007), and few studies have documented both genetic and fitness consequences of decreased immigration through time in a natural population.

We investigated temporal changes in immigration rate and inbreeding in a long-studied population of Florida Scrub-Jays (*Aphelocoma coerulescens*; hereafter FSJ). In the past century, FSJs have faced drastic population declines caused by human-mediated habitat loss and destruction, and the geographic distribution of the FSJ is now highly fragmented across its ancestral range (Woolfenden and Fitzpatrick 1996). The FSJ population in the immediate area of our study site has remained stable because of outstanding habitat management of the local environment. However, our study population is nested within a larger metapopulation (Coulon *et al.* 2008; Stith *et al.* 1996). The surrounding regional population continues to undergo drastic declines in available habitat and numbers of jays, and extirpations of small subpopulations have been documented in the past decade. In 1992-1993, a comprehensive survey of the FSJ concluded that the FSJ had declined in number by 25-50% in the past decade alone (Fitzpatrick *et al.* 1994). A recent statewide survey in 2010 estimates a minimum decline of 35-40% since the 1992-1993 survey (Boughton and Bowman 2011).

Archived blood samples of FSJs in our population from 1988 to the present

allow us to test whether recent regional population declines have resulted in decreased immigration rates and a subsequent loss of genetic diversity despite the demographic stability of our study population. Here, we examine changes in immigration rate, heterozygosity, and inbreeding over a 23-year period. As hatching failure is a common consequence of inbreeding in birds (Bensch *et al.* 1994), we also investigate the relationship between parental relatedness and hatching failure.

4.3 Methods

4.3.1 Study population

The FSJ is a Federally Threatened bird restricted to the xeric scrub habitat unique to the Florida peninsula. A population of FSJs at Archbold Biological Station (Highlands County, FL) has been intensively studied since 1969 (Woolfenden and Fitzpatrick 1984; Woolfenden and Fitzpatrick 1991). Every individual in the study population has been uniquely banded, allowing documentation of new immigrants each year. Because FSJs are genetically monogamous (Quinn *et al.* 1999; Townsend *et al.* 2011), accurate pedigrees can be constructed from field observations alone. All nests of all family-groups are monitored (clutch sizes, nestlings, and fledglings), producing fully documented annual fecundity and fitness measures for all breeding birds. We have blood samples for every nestling and immigrant recruited into the study population in 1989-1991, 1995, and from 1999 to the present.

4.3.2 SNP discovery and Beadchip design

A set of genome-wide single nucleotide polymorphisms (SNPs) was discovered in the FSJ using genotyping-by-sequencing (GBS) of 103 individuals in 27 nuclear families from 1989 and 2008 (Chen *et al.* 2014). We sampled individuals from these two different time periods to help minimize ascertainment bias. SNPs were called using both a custom reference-free pipeline (described in Chen *et al.* 2014) and a reference-based pipeline. For the reference-based pipeline, demultiplexed and adapter-trimmed reads were aligned to the draft FSJ genome (Chen *et al.* in prep) using BWA (Li and Durbin 2009). We used Picard tools (http://picard.sourceforge.net) to sort and merge the individual BAM files before indel realignment and variant calling using GATK (DePristo *et al.* 2011). Because the length of flanking sequence required for Illumina iSelect Beadchip assays (50-60 bp on either side) is greater than the GBS read lengths, we could only use SNPs that could be aligned to the FSJ genome. SNPs called using the reference-free pipeline were mapped to the FSJ genome using BWA, and the two sets of SNP calls combined based on their physical location. After thinning to one SNP per 100 bp window, there were 41,853 SNPs. Summary statistics for all SNPs were calculated using custom Perl scripts or VCFtools (Danecek *et al.* 2011), and a quantitative score of Mendelian inheritance was assigned using MendelChecker (Chen *et al.* 2014).

We used a number of criteria when designing custom Illumina iSelect Beadchips. First, we filtered out all sites with low mapping quality or read depth (MQ<35, QD<2), high levels of missing data (>8%), excess heterozygosity (>75%), low minor allele frequency (MAF<0.02), or low Mendelian inheritance scores (M<-20). We removed SNPs that were fewer than 50 bp from the end of

a scaffold or had more than 2 alleles. Flanking sequences for each SNP assay were derived from the draft FSJ genome. We checked for repetitive elements in the flanking sequences using RepeatMasker (http://www.repeatmasker.org) and removed any sites near repetitive elements. The remaining 19,087 SNPs were submitted to Illumina's Assay Design Tool for evaluation. Each SNP was assigned a score that represents the expected success rate of the assay. The final 20k Beadchip design consisted of 17,628 SNPs, each with a minimum score of 0.781 and MAF>0.0223. The number of SNPs remaining after each filtering step are shown in Table C.1.

Custom iSelect Beadchips were manufactured by Illumina, and assay design was successful for 15,416 SNPs. Putative chromosomal locations of SNPs were assigned by aligning the FSJ genomic scaffolds to the Zebra Finch genome using standalone BLAST (Camacho *et al.* 2009) and picking the best BLAST hit. Given the high degree of synteny among extant bird lineages (Ellegren 2010), we are confident that these SNPs are distributed across the genome (Figure 4.1).

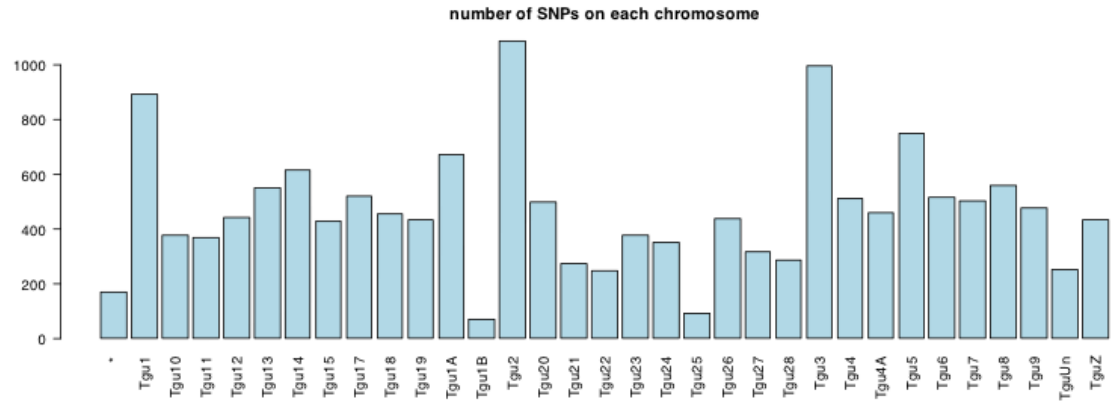

Figure 4.1: Putative chromosome locations of our genome-wide SNPs. Putative genomic location was determined by aligning reads to FSJ genome scaffolds, which were aligned to the Zebra Finch genome. Here we show the number of SNPs aligned to each Zebra Finch chromosome. TguUn contains scaffolds that were not successfully assigned to a chromosome.

4.3.3 Genotyping

Historically, genomic DNA has been extracted from blood samples stored in lysis buffer using a variety of methods, including phenol-chloroform protocols and Qiagen or E-Z DNA extraction kits. For the archived DNA samples that were low in concentration, we re-extracted genomic DNA from blood samples using the Qiagen DNeasy kit. DNA samples from different years were mixed in a semi-randomized order on 96-well plates to minimize any batch effects. A total of 4,032 samples was genotyped with the custom BeadChips at Geneseek, Inc. (Lincoln, NE), representing 3,984 unique individuals. For positive controls, 1 individual was genotyped 42 times and 7 individuals were genotyped twice.

Genotyping results were analyzed using GenomeStudio (Illumina, San Diego). After excluding SNPs with Gentrain scores < 0.7 and call rates < 90% as well as samples with call rate < 95%, a set of 14,151 SNPs in 3,770 individuals was exported to PLINK format. Reproducibility between duplicate samples was high (>98%). We checked for Mendelian inconsistencies using PLINK v1.07 (Purcell *et al.* 2007) and PedCheck (O'Connell and Weeks 1998), and removed 192 individuals and 99 SNPs with high Mendelian error rates. To obtain unbiased estimates of genetic diversity and relatedness in this study, we wanted only autosomal SNPs in approximate linkage equilibrium. We excluded 365 SNPs on the Z chromosome because sex-linked SNPs would skew our estimates of mean heterozygosity. We pruned SNPs in high linkage disequilibrium (LD) using the PLINK option –indep 50 5 2. Our final cleaned and LD-pruned genotype data set consisted of 7,404 autosomal SNPs in approximate linkage equilibrium in 3,578 individuals. We have near-complete sampling of all nestlings from 1991, 1995, and 1999-2013 and all breeders from 1990-1991 and 2003-2013 (Figure 4.2).

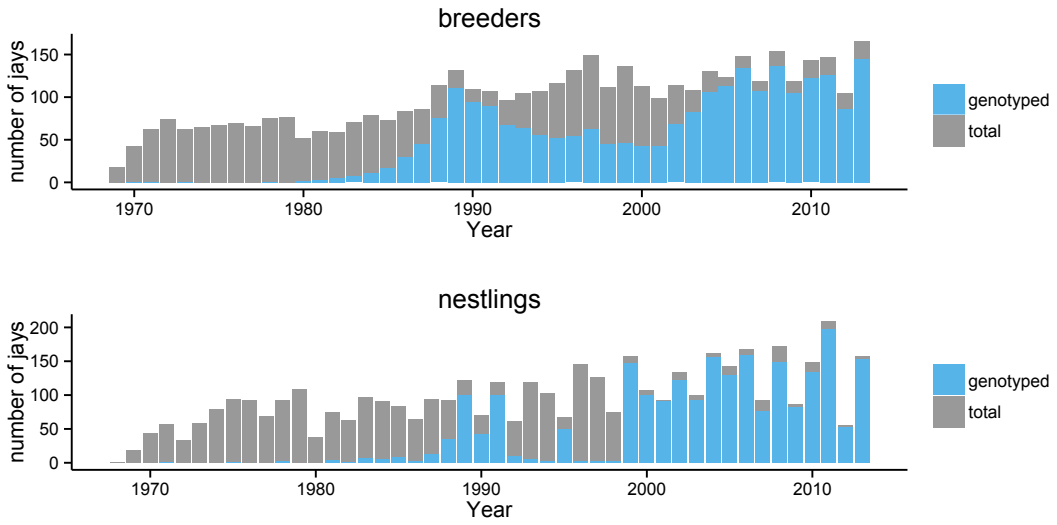

Figure 4.2: The number of breeding adults and nestlings born in the study tract each year from 1990-2013. Gray bars indicate the total number of individuals, and blue bars indicate the number of individuals that were genotyped. We have genotypes for >75% of all breeders and nestlings in 1991 and 2003-2013.

4.3.4 Population genetic analyses

Because the area of the study tract and therefore the number of monitored FSJ territories were purposely expanded during the late 1980s, we restrict our analyses to a core set of 48-77 territories from 1990 onwards. We classify a breeder as an immigrant if it was known to be born outside the core tract. The genealogical relationships obtained from field observations were checked for consistency and separated into component pedigrees using the program MORGAN (Thompson 1994). Pedigree-based inbreeding coefficients were calculated for each individual using the program PedigreeViewer. For each breeder and nestling in each year for which we had >50 individuals genotyped, we estimated individual inbreeding coefficients from our genomic data using PLINK (option –het). Mean site-based observed heterozygosity for each individual was calculated as the number of heterozygous loci divided by the total number of loci genotyped in that individual. Mean pairwise Identity-by-Descent (IBD) between all possible male-female pairs or mated pairs for each year was calculated with the PLINK option –genome. A breeding pair was classified as "immigrant" if at least one of the breeders was born outside the study tract and "resident" if both breeders were born within the study tract. All statistical analyses were performed in the R statistical package (R Core Team 2013).

4.3.5 Correlations with fitness components

We tested for a relationship between pairwise IBD of mated pairs and two proxies for fitness: clutch size and hatch failure. We obtained clutch size and hatching success from 604 nests with genotyped parents from 1990-2013. An

egg was considered a hatching failure if it remained unpipped more than 5 days after the other eggs had hatched. Nests that were depredated or abandoned less than three days after the eggs hatched were excluded from this analysis. We only included the first clutch that hatched for each pair within a year. From the long-term demographic data, we obtained measures of breeding density per year, the breeding age of the female, and the date each clutch was completed. As the area of the core study tract remained approximately constant through time, we used the number of territories each year as a proxy for breeding density. We determined whether a given pair had previously bred together, versus the nest being their first breeding attempt. Pairwise IBD of each breeding pair was calculated from the genomic data using PLINK. Daily rainfall in inches and drought index data were obtained from the Archbold Biological Station weather station (http://www.archbold-station.org/station/html/datapub/data/data.html). Drought index is a number from 0 to 800, with 0 indicating no drought and 800 indicating maximum drought. We considered mean rainfall and drought index in the breeding season (March-May).

We used generalized linear mixed models as implemented in glmer from the lme4 package in R (Bates *et al.* 2004). We fit logistic regression models for hatching failure, coding the number of unhatched eggs as the number of failures and the number of hatched eggs as the number of successes. Any eggs that were depredated before hatch were not counted. Clutch size was analyzed using linear mixed models. For each dependent variable, we first determined which independent variables (density, rainfall, drought index, pair experience, female age, lay date, and pairwise IBD value) were important by fitting models for each independent variable separately as a fixed effect and the identity of

the pair (coded as the male and female US Fish & Wildlife Service numbers) as a random effect. Then, we constructed models for all combinations of significant predictors and performed model selection using the corrected second order Akaike information criterion, AICc, which takes into account sample size.

4.4 Results

4.4.1 Decreased migration through time

The number of new immigrants arriving in our study population ranged from 3 to 24 each year, and the proportion of breeders that are new immigrants ranged from 0.02 to 0.21. The proportion of total immigrant breeders in any given year ranged from 0.29 to 0.58. Both the proportion of new immigrants (adjusted R^2 = 0.4211, p = 0.0003606) and the proportion of total immigrants (adjusted R^2 = 0.85, p = 9.577e-11) in the breeding population declined during the course of our study period (Figure 4.3).

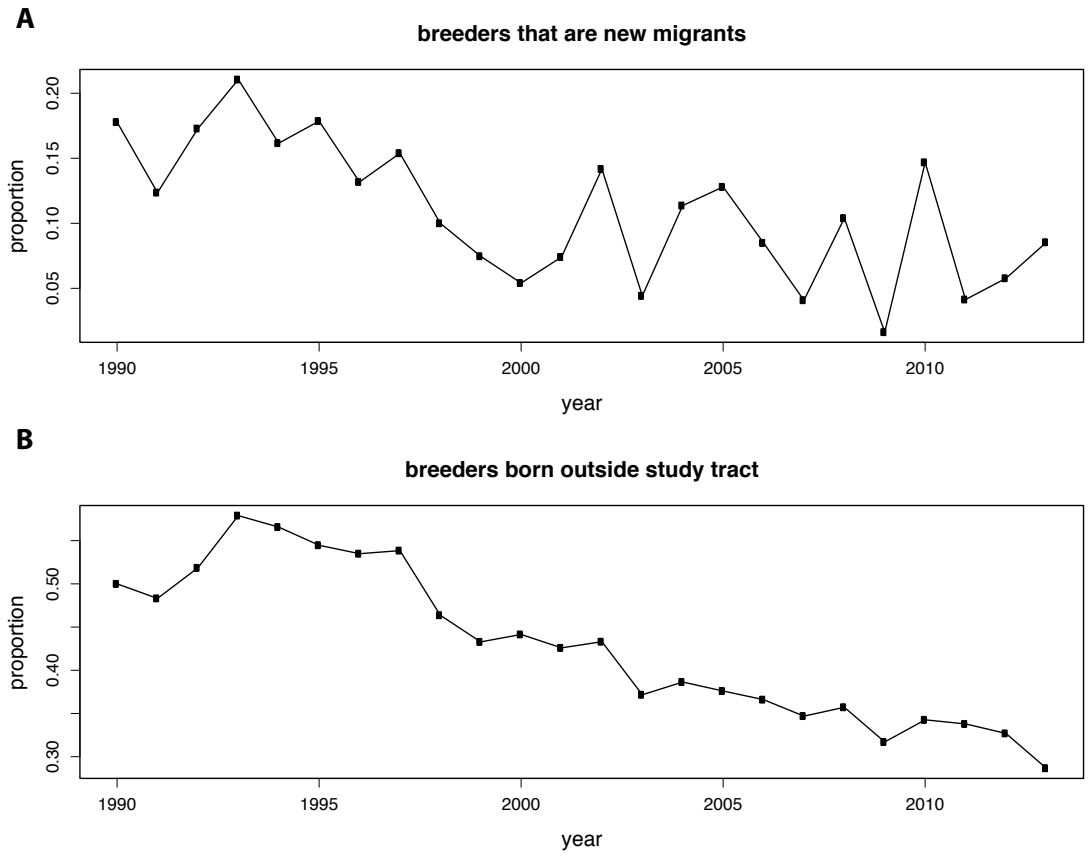

Figure 4.3: Immigration into Archbold is decreasing. (A) The proportion of breeders that are new immigrants each year is declining (adjusted R^2 = 0. 4211, p = 0.0003606). **(B)** The proportion of breeders that were immigrants each year is declining (adjusted R^2 = 0.85, p = 9.577e-11).

4.4.2 Population genetic consequences of immigration

We genotyped 3,578 individuals in our study population through time at 7,404 autosomal SNPs in approximate linkage equilibrium. To investigate the genetic contribution of immigrants to the population, we used PLINK to estimate individual inbreeding coefficients and mean observed heterozygosity as well as pairwise IBD values. We compared inbreeding coefficients estimated from the genomic data with those estimated from the pedigree, and the two measures were significantly correlated (Pearson's r = 0.733982, p < 2.2e-16). Because the pedigree-based estimates are actually expected inbreeding coefficients, we include only SNP-based estimates in the analyses below.

We used linear regressions to test for directional changes in inbreeding levels or mean pairwise IBD sharing through time. Immigrant breeders had significantly lower levels of observed heterozygosity compared to resident breeders (Wilcoxon rank sum test, p < 2.2e-16; Figure 4.4). Breeding pairs with at least one immigrant had lower IBD sharing than breeding pairs with two residents (Wilcoxon rank sum test, p = 0.0001949; Figure 4.5). Genetic similarity of all breeders, as measured by the mean pairwise IBD between all possible pairwise male-female combinations, increased through time (adjusted R^2 = 0. 3166, p = 0.00577). Pairwise IBD between all observed pairs increased with time (adjusted R^2 = 0.4819, p = 0.000412).

Mean inbreeding of the birth cohort is negatively correlated with the proportion of breeders that are immigrants (adjusted R^2 = 0.3013, p = 0.01318; Figure 4.6A). Linear regression analysis shows that mean inbreeding of the birth cohort increased through time (adjusted R^2= 0.3019, p=0.01308; Figure 4.6B). In 2009-2012, nestlings with two resident parents had significantly higher

mean inbreeding coefficients than nestlings with at least one immigrant parent (Wilcoxon rank sum tests, $p < 0.05$).

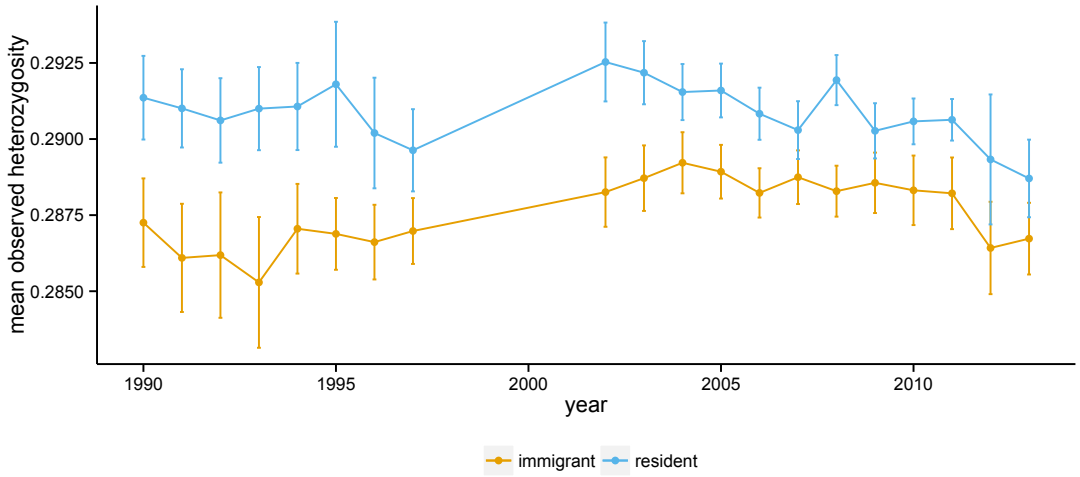

Figure 4.4: Mean genome-wide observed heterozygosity for immigrant and resident breeders from 1990-2013. Only years with more than 50 genotyped breeders are included. Immigrant breeders have lower observed heterozygosity compared to resident breeders ($p <$ 2.2e-16).

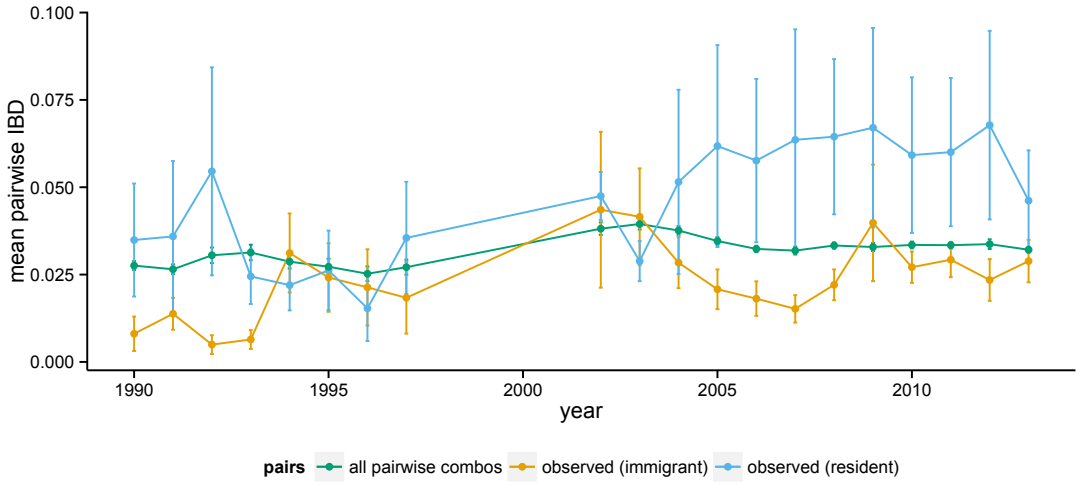

Figure 4.5: Relatedness among breeding pairs through time. This figure shows the proportion of IBD sharing between all possible male-female pairs (green), observed immigrant pairs (yellow), and observed resident pairs (blue) for each year with more than 50 genotyped breeders. Breeding pairs with at least one migrant are less related compared to pairs consisting of two residents (p = 0.0001949). In recent years, the proportion of IBD sharing between resident breeders is higher than expected.

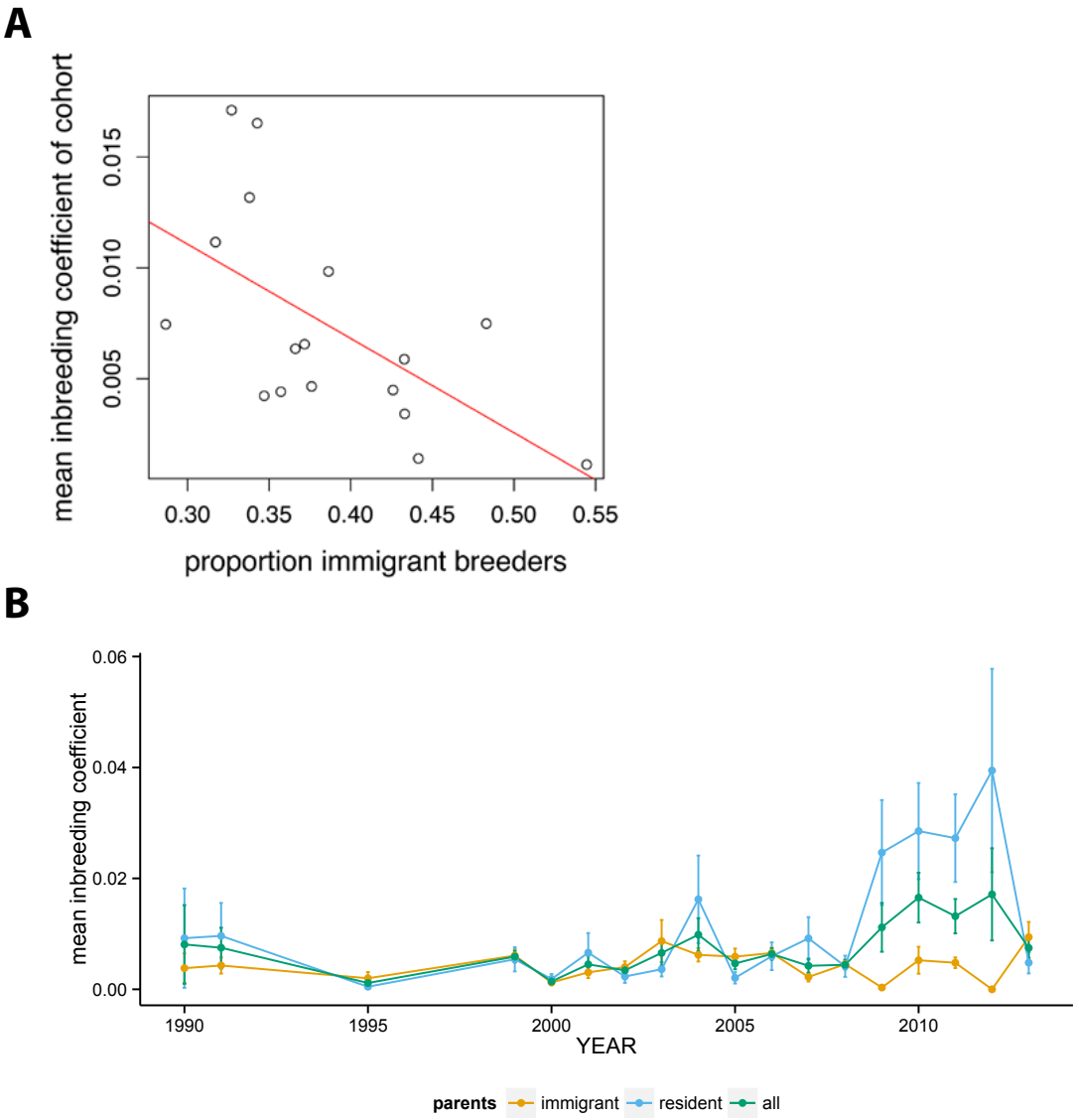

Figure 4.6: Decreasing immigration is correlated with increasing inbreeding. **(A)** Relationship between the proportion of immigrant breeders in the population and mean inbreeding coefficient of the birth cohort. Mean inbreeding of the nestlings born in a given year is negatively correlated with the proportion of breeders that are immigrants (adjusted R^2 = 0.3013, p = 0.01318). **(B)** Mean inbreeding coefficient of all nestlings (green), nestlings with at least one immigrant parent (yellow), and nestlings with two resident parents (blue) for each year. Overall, mean inbreeding coefficient in the birth cohort is increasing through time (adjusted R^2 = 0.3019, p = 0.01308). Nestlings with immigrant parents are less inbred than nestlings with resident parents in 2009-2012.

4.4.3 Fitness consequences of increased IBD

We assessed the relationship between pairwise IBD sharing between mated pairs and hatch failure or clutch size. Of the 603 nests we considered, 150 had at least one egg that failed to hatch (33%). Out of 2,034 eggs, 184 (9%) failed to hatch. The proportion of eggs that failed to hatch in a given year varied from 0 to 0.13. The only independent variable that significantly correlated with hatch failure was pairwise IBD of the parents (Figure 4.7). Clutch size varied from 2 to 5, with a mean of 3.5. For clutch size, lay date, breeding density, drought index, breeding experience, and female age were all significant predictors, but the only variables shared across the top four models were lay date and density (Table 4.1). We found no genetic correlations with clutch size.

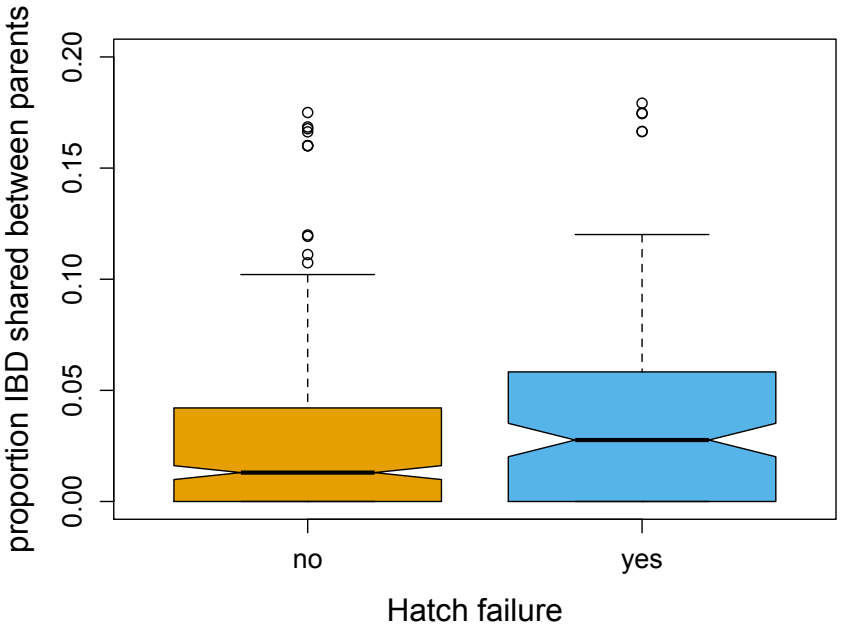

Figure 4.7: Relatedness of parents predicts hatching failure. The breeding pair for nests with at least one unhatched egg has higher IBD values than the breeding pair for nests with no hatch failure.

Table 4.1: Model results from an analysis of factors involved in explaining clutch size in the FSJ from 1990-2013. Here, df is the number of parameters in the model, AICc is the AIC value corrected for small sample size, and ΔAICc is the difference between the AICc value of the given model and that of the best model.

Model	**df**	**AICc**	**ΔAICc**
LayDate + Density + DroughtIndex + Experience	5	1192.728	0
LayDate + Density + Experience + femaleAge	6	1193.985	1.257
LayDate + Density + femaleAge	6	1196.304	3.576
LayDate + Density + DroughtIndex + femaleAge	7	1197.35	4.622
LayDate	4	1198.687	5.959
LayDate + DroughtIndex	6	1200.222	7.494
LayDate + Experience	5	1200.608	7.88
LayDate + Density	5	1201.231	8.503
LayDate + Density + DroughtIndex + Experience + femaleAge	7	1201.545	8.817
LayDate + DroughtIndex + Experience	5	1202.056	9.328
LayDate + femaleAge	6	1202.885	10.157
LayDate + Density + DroughtIndex	6	1204.398	11.67
LayDate + DroughtIndex + Experience + femaleAge	7	1204.465	11.737
LayDate + Density + Experience	8	1205.553	12.825
LayDate + Experience + femaleAge	6	1205.723	12.995
LayDate + DroughtIndex + femaleAge	7	1207.798	15.07
Density + Experience	5	1214.875	22.147
Experience	4	1215.527	22.799
Density	4	1216.92	24.192
Density + DroughtIndex + Experience	6	1219.388	26.66
DroughtIndex + Experience	5	1220.568	27.84
femaleAge	4	1220.587	27.859
Density + femaleAge	5	1220.827	28.099
Experience + femaleAge	5	1221.38	28.652
Density + Experience + femaleAge	6	1222.042	29.314
Density + DroughtIndex	5	1222.266	29.538
DroughtIndex	4	1225.144	32.416
Density + DroughtIndex + femaleAge	6	1225.945	33.217
DroughtIndex + femaleAge	5	1226.266	33.538
DroughtIndex + Experience + femaleAge	6	1226.444	33.716
Density + DroughtIndex + Experience + femaleAge	7	1226.561	33.833

4.5 Discussion

Our study population has remained approximately stable in number for decades, yet its proportion of both new and total immigrant breeders has declined through time, presumably reflecting regional habitat loss over this period. The surprisingly high proportion of breeders that are immigrants indicates that our study population is nested within a larger metapopulation. We found that immigrants were significantly less heterozygous compared to residents, and we suggest two possible explanations for this pattern that are not mutually exclusive. First, immigrants to the Archbold population could tend to include individuals originating from smaller, more isolated, and presumably more inbred populations. Second, individuals who disperse from other populations could have lower heterozygosity compared to individuals who do not disperse because of some behavioral correlation. Thorough sampling of regional populations, and closer scrutiny of individual behavior and heterozygosity levels within our population, will perhaps help disentangle these two hypotheses. Regardless of the causes of lower genetic diversity within immigrant individuals, our results provide evidence for the importance of immigrants in contributing novel genetic variation to the population over time. Specifically, immigrant-resident pairs have lower than expected IBD sharing given overall levels of relatedness among all breeders, and we find evidence of inbreeding in resident-only pairs. Mean IBD sharing between pairs and mean inbreeding coefficient of the birth cohort have been increasing with time.

Preliminary analyses of our study population show that inbred individuals had significantly lower reproductive success (measured by total fledglings produced) than did the pool of outbred individuals (Mann-Whitney U, $p = 0.004$;

Chen *et al.* unpubl. data). Negative fitness consequences of inbreeding should lead to the evolution of inbreeding avoidance mechanisms, which could happen via disassortative mate choice based on relatedness or dispersal (Pusey and Wolf 1996). Levels of IBD sharing between observed resident-resident breeding pairs is higher than expected in most years, whereas IBD sharing observed between resident-immigrant pairs is lower than expected. These results suggest that there likely is no mate choice based on genetic similarity, but there could be inbreeding avoidance via natal dispersal in this population. Direct tests for assortative mating could be carried out using the available genotype and demographic data. This analysis is in progress.

Mean inbreeding coefficient for nestlings with resident parents increased from 0.004 in 2008 to 0.025 in 2009. During the autumn of 2008, the study population experienced one of the highest recorded monthly breeder mortality rates since 1969 (Figure 4.8). Screening of 76 FSJ blood samples revealed 75% prevalence of Eastern Equine Encephalitis antibodies, and birds that died had poor body condition (Wilcoxen *et al.* 2010). That year, environmental conditions were suitable for high mosquito densities, and both high levels of local encephalitis and no increases in densities of known predators were documented, suggesting a possible role of disease in the high mortality event. We speculate that the significant increase in inbreeding coefficient for offspring of resident breeders could be an ephemeral signature of the bottleneck event, similar to that observed in Song Sparrows (Keller *et al.* 2001). However, more work is needed to fully test this hypothesis. There have been three other high mortality events in our population (in 1979, 1989, and 1997), but unfortunately two of these events occurred during the early study tract expansion phase, and we do not have sufficient sampling of the population in 1997.

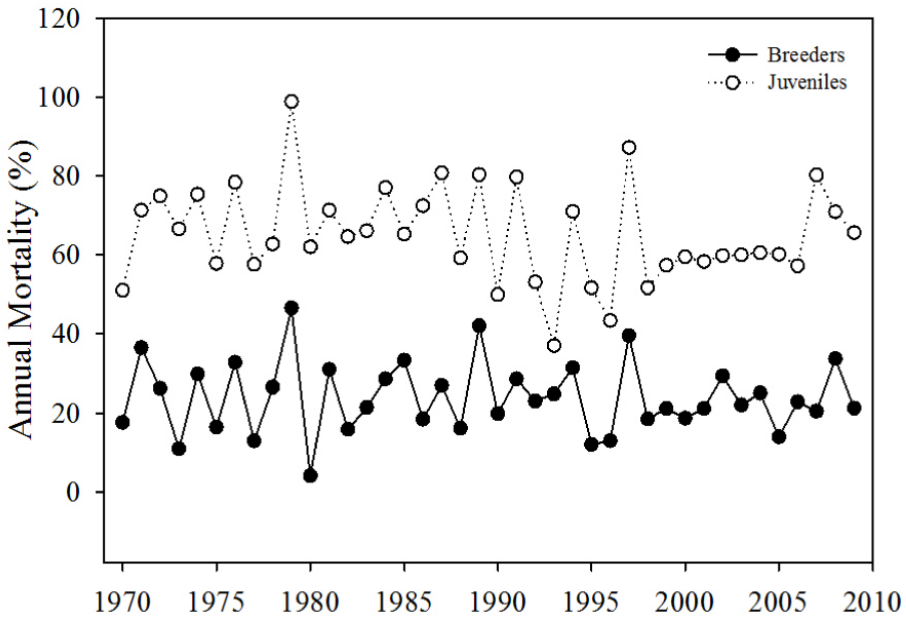

Figure 4.8: Annual mortality of breeders and juveniles at Archbold from 1970-2010. Over this 40-year time span, acute high mortality events putatively caused by encephalitis disease occurred in 1979, 1989, 1997, and 2008.

The association between decreased hatching success and increased IBD sharing between parents replicates similar findings in other bird species (Bensch *et al.* 1994; Kempenaers *et al.* 1996; Hansson 2004). Hansson (2004) investigated marker-based relatedness of parents with additional measures of reproductive success in Great Reed Warblers and also found no association of relatedness with clutch size. In the nests we studied, other environmental variables known to be correlated with hatching failure in birds were not significant predictors. This result is surprising given that a previous study investigating hatching failure in an adjacent study population of FSJs found that hatching failure was associated with rainfall during the breeding season, pair experience, and female age (Wilcoxen *et al.* 2011). That study did not consider genetic relatedness between mated pairs and used a slightly different modeling scheme, but we still expected to find at least some overlap in contributing factors. We found lay date, density, drought index, pair experience, and female age were significant predictors for clutch size, but not hatch failure. However, an important caveat is that we did not include all possible factors that could explain clutch size or hatch failure, and results may be different after inclusion of additional relevant variables (*e.g.*, presence of helpers, rainfall the previous winter, etc.). Here, we do not distinguish between the two possible causes of hatch failure (infertility or early embryo mortality), but a study in Zebra Finches suggests that early embryo death is the more likely outcome of inbreeding depression (Hemmings *et al.* 2012). Future work will investigate the impact of inbreeding on other components of fitness.

We used a conservative definition of immigrants in this study. Birds that were born in the South tract, another long-studied set of territories immediately adjacent to our study tract, were classified as immigrants. Even though we

included birds from elsewhere in the same population, we still find evidence for lower IBD sharing between immigrants and residents. South tract birds are likely to be less related to our core study tract birds because FSJs have short natal dispersal distances (Woolfenden and Fitzpatrick 1984), and we have preliminary evidence of isolation-by-distance within the greater Archbold population. However, future work should consider the South tract birds separately to more accurately measure the genetic contribution of birds from different populations in the region.

In this paper, we focused on SNP-based estimates of inbreeding and IBD sharing because our pedigree, despite being one of the most extensive pedigrees for any wild species, does not completely capture all relationships between individuals in the wild. Any kinship coefficient calculated from the pedigree for a pair containing a migrant will be zero even if the true kinship coefficient is higher than zero. However, because we do not have genotypes for every individual in our population through time, pedigree-based analyses would provide a greater sample size. In the future, it will be important to repeat the analyses above with pedigree-based inbreeding and kinship coefficients and compare results to those obtained with the SNP-based estimates. In addition, we can boost our sample size by using the pedigree and available SNPs to impute genotypes for any individuals that lack genotype data.

IBD estimates in this study were obtained using PLINK's single-marker IBD approach, but future work will use more sophisticated programs (*e.g.*, BEAGLE; Browning and Browning 2011; ALADIN; Albers *et al.* 2008) to generate more accurate estimates of genome-wide IBD sharing between pairs of individuals. We are in the process of generating a dense linkage map, which will allow

us to infer haplotype blocks in the founders and trace the descent of these genomic segments down our pedigree. We will compare the frequencies of these blocks in each generation to distributions obtained from gene-dropping simulations to assess the impact of natural selection and gene flow in maintaining genetic variation in the population over time. In addition, this study focused on genome-wide estimates, but the same questions are currently being asked at the haplotype level. We will trace the fitness impact of novel immigrant alleles to determine the net importance of immigrants on population genetic diversity. This analysis will test whether there is a decline in novel genetic variation being introduced by immigration in different genomic regions.

Here we showed that decreased levels of immigration have led to increased inbreeding in the Archbold population of FSJ. Previous work has shown that effective dispersal of FSJs decreases as habitat fragmentation increases (Coulon *et al.* 2010). Our results inform conservation management decisions for the FSJ by placing additional emphasis on the importance of preserving habitat in a landscape configuration that maintains dispersal. In particular, this study demonstrates that even small and perhaps inbred populations may play a vital role in preserving genetic diversity in larger and seemingly stable populations. We suggest that conservation efforts need to pay particularly close attention to local population declines, as there can be strong departure from homogeneity across even nearby populations. This study is an example of how combining genomic data with demographic and pedigree data from one of the longest-studied endangered species is a powerful approach to addressing fundamental questions concerning the population genetic consequences of declining population size.

4.6 Acknowledgments

We thank Jen Grenier, Charlotte Acharya, and Laura Stenzler for help with sample organization and labwork. Cris Van Hout, Haley Hunter-Zinck, and Angela Early provided statistical advice. None of this work would be possible without many years of hard work by John Fitzpatrick, Reed Bowman, Raoul Boughton, Shane Pruett, and many students, interns, and staff at Archbold Biological Station. Thanks to the Clark and Harrison labs for comments. This work was supported by NSF (SGER DEB 0855879 and DEB 1257628), the Cornell Lab of Ornithology Athena Fund, the Andrew W. Mellon Student Research Award, and the EEB Graduate Student Research Fund. N.C. was supported by a NSF Graduate Research Fellowship and a Cornell Center for Comparative and Population Genomics Fellowship.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Table A.1: List of W candidate contigs tested by PCR. Note that contigs that do not have female-specific markers may still be located on the W chromosome.

Contig name	Female-specific marker?	Left primer	Right primer	Annealing temp
Contig197.14	N	TAGGAATGGGATTCAGCCAG	TGGAAAACCGCATTGTGTAA	60 °C
Contig197.23	N	ACCCTAAGCCTGACGGATCT	TTCCCATAATGTCACCCCAT	60 °C
Contig212.15	N	AATGAGGTCTCACACCAGGG	TAGACACCACGCAAAAGCAG	60 °C
Contig212.18	N	ACCTGAAGCAGAGTGGCATT	AGGACAGTGTGTCCCCTCAC	60 °C
Contig212.8	N	ACTGCTGCTTTGATTGGCTT	AAGGGTGCTTCATTGTCTGG	60 °C
Contig230.11	N	CTGGGTCCATACCTGCACTT	GTCACATTTAGCCAGGCCAT	60 °C
Contig243.11	N	GCACATACCCACAACTGCAC	GCGCAAAGTACGCCTCTTAC	60 °C
Contig245.12	N	TTGCACATGGATGAGAGTCC	AGGCGTACAGCATCATTCCT	60 °C
Contig248.17	N	AGCTTGCACCCTGTCTGAGT	GGGAAACAAGGAAGCTAGGG	60 °C
Contig256.6	N	GCTGGCACTCAGGTAAAAGC	ACCTAACGAGCTCAGCAAGC	60 °C
Contig260.11	N	GAACCTGTTCCCTAGCACGA	GGCATGTATTCTGGGCAAGT	60 °C
Contig269.4	Y	TTACACCTTGAAGGGCTTGC	TTGAGGCTTCATTACTGGGG	60 °C
Contig276.6	N	TCGAAATGGTGCTACACAGC	CCCTCAAAATACCTCTGGCA	60 °C
Contig282.24	N	GACCTCTGCTTCCTCCTTTG	TGGGACTCTCTGTTGGTTTG	59 °C
Contig284.11	N	CCAACTCTAGCTGCGAATCC	TAACAGGTCCTGGAGATGGG	60 °C
Contig286.6	N	CTTAAAGCTTGGGAGGAGGG	GTCTGCGAGAGGGGAATAAA	60 °C
Contig287.19	N	ACTGCGACCTTTGAAACCAC	AGCTACGCTCGTGTCTTCGT	60 °C
Contig291.5	N	TTCCGTAGCGTTTGTTCCTT	ATAGTGACGTTGGCTTTGGG	60 °C
Contig297.8	N	ACTCAAGCCAGAGGCGATAA	AGGACAGTGTGTCCCCTCAC	60 °C
Contig303.13	N	GCAGGCTGGGTTAGTCAGAG	GTGTGGCAGATCACAGGAGA	60 °C
Contig317.11	N	TGTTGTGCCTCATTCCTTTG	GATTTCTGCTGCTCGCTTG	60 °C
Contig330.15	N	GCATTGATTGGTCTTGCTCA	GGAATCCTACAGGAACGCAA	60 °C
Contig337.3	N	ACAATGGCGAGAGCTCAACT	TACAAAAACCTTTCGGTGGC	60 °C
Contig337.4	N	GATGAAATGCAAAGGCTCGT	TGAGATCCTCCTGGCTATGG	60 °C
Contig339.4	N	ATGCTGGTTTTCCCAGTGAG	TACAGGGTTTCCCGTGCTAC	60 °C
Contig350.1	N	GGAGCCTGACACGTCTTTTC	CTCCGCATAGTATTTCCCGA	60 °C
Contig350.3	N	CAGCAAGATGTGGCAAAGAA	TGGGCAGGAGTAGATGGTTC	60 °C
Contig353.5	Y	GCTGGGACCACAAGTTCATT	TGCCGACATTTCACATCCTA	60 °C

Continued on next page

Table A.1 – *Continued from previous page*

Contig name	Female-specific marker?	Left primer	Right primer	Annealing temp
Contig353.6	Y	GTGTTGATCCCTCTGCCATT	GGAAAAGATCCCCCACAAAC	60 °C
Contig360.2	N	CAATGGTAGGCTGTCCTGGT	ACAGCGGGTGGTACTTTCAC	60 °C
Contig370.2	Y	TGTCTCCAAAGCAACAGACG	GCAGTTGCAATCTGGTGTGT	60 °C
Contig379.5	Y	CCAACGCCTCATACGTAAAAA	TGAAGCAGGTAGTTGGGTTG	60 °C
Contig380.10	Y	CCGTGTTATAACCCCACTCG	GCAGGGGGTTGGAATTTAAG	60 °C
Contig380.12	N	CCCCTTTGCAATCAGAACAT	AACAATTGGGACCACAAAGC	60 °C
Contig380.6	Y	AAATCCTTGGACGCCTCTCT	CCACGTGTGCATCTATGGAG	60 °C
Contig395.5	Y	CAATTGTCTTCCTTCGGGAG	TCAGCATCCTCAGACACGAC	60 °C
Contig395.6	Y	TGATTCATGCCGATTTTTCA	TTCCGAGATCTTTTGAAGGC	60 °C
Contig395.7	Y	GGGTTAAACGCACCTCATTC	TACCCCACGATTCAAAAAGC	60 °C
Contig399.2	N	CTCCTGAGTCCCACGGATAA	TTCTCTGCATGTGTTGAGGG	60 °C
Contig401.2	N	TCAATGGGGTAATAGCAGGC	GTGCTGGAAGTGGACGGTAT	60 °C
Contig414.2	N	CCACCGTTTTGTTTCGAGAT	TTGATCCCCCTTAAGAGCCT	60 °C
Contig415.2	N	TCAGGGATGAGAGCGTTTCT	GGCTCCACATGACCCATAAC	60 °C
Contig416.1	N	GCTCTTCTGAGGACGGACAC	CTCCATCACCACACAAAACG	60 °C
Contig419.6	N	TGTCAGTAGGAGAGTAGGAAGTAAGG	CTGGGTCAATGCTGTTCTTG	59 °C
Contig432.8	Y	GCAGGGGATTATGTCCTCAC	CTGAGGGTTCCTCATGCAAT	60 °C
Contig439.1	N	ACCGGCTGGAAATTTTCTCT	AGATGCAAAATCTGTGCCCT	60 °C
Contig451.2	N	CTAGCTCACCAAAACCAGCC	AGCTCCCAAGAGACTGACCA	60 °C
Contig455.5	Y	GAGAAGAATCTGCATTGGTGC	TCAACTTGATGAGGGTTCAGTA	60 °C
Contig457.3	N	GTAGTGGTCGACCTTCCCAA	ATTAGGCCGTGTGTGTCTCC	60 °C
Contig463.1	Y	CCACCTGCCAACCTATCAGT	GCAGAAATTGTCGGTGTCCT	60 °C
Contig473.3	Y	TACCTGGAAACCCATACCGA	TGTGCTTCTTGTGCCCTACA	60 °C
Contig508.1	N	GCGTGTTGTTGTATGTTGGG	ACTCGTGACCGTAGGCAAAT	60 °C
Contig512.3	N	TGCTTTCACAAGTGATTCGG	AGCGCAGAGCTACCTGAAAG	60 °C
Contig516.1	Y	CTCCCTGCCCAGATAAATCA	TAGGCATTCCAACGACTTCC	60 °C
Contig521.3	N	AGGCCTTGCTTCAAGGTACA	TGATACATCAGGCTTTGGCA	60 °C
Contig522.2	Y	ATCTTCGGTCAAGTTGGGTG	CTCCCAAATTTGTGGGTGTT	60 °C
Contig522.5	Y	TGGGAGTGCATGTTGTGAAT	CACATGCTCCAGGCATTAGA	60 °C
Contig522.6	N	CCTGAATCCATCCCTGAATC	ACAATGCCTTGGACGTGAAT	60 °C

Continued on next page

Table A.1 – *Continued from previous page*

Contig name	Female-specific marker?	Left primer	Right primer	Annealing temp
Contig528.3	N	ATCAGCCCACCTTAGGGAGT	CCCATTCAAGGACTAGGCAA	60 °C
Contig529.1	N	TCCTGCTGATGTCCCACATA	CTTCCCAACCACAAGAGCAT	60 °C
Contig558.2	Y	GAAAGACGCAAAATGCAGGT	GCACCTTGGAACAATCAGGT	60 °C
Contig563.3	N	AACCCCAGTCACACATAGCC	CACTTGGCTGACCCATACCT	60 °C
Contig596.1	N	CGCATTAGGATTTCTTCGGA	CTGCTGCCTACGTGTTGAGA	60 °C
Contig611.3	N	TGAGACCCGAATCCCATAGA	TGCACCTAGATCTGTGCCTG	60 °C
Contig631.3	N	ATGCTATCCTGCCCAATCAC	GGACCGTATGATGGTGGAAC	60 °C
Contig639.1	N	GATACGTGCCCCTTCAGAAA	TTCAGACGTGTAGTGCGTCC	60 °C
Contig659.3	N	TGTCGCTACAGCTGTTTTGG	CACTGCAACCGTATGTGAGC	60 °C
Contig677.1	N	CTCTGCTCAGTGAACTCCCC	TTCAGCCTTCAAATTACCCG	60 °C
Contig684.2	N	AACCAATGGCAAAAGAGACG	CGGTGAGAAGCCGCTAATAA	60 °C
Contig697.3	N	CCCTTTTACCACGAGTTCCA	GAATGCATTGGGCACAGAG	60 °C
Contig698.1	N	CCATCTCAAGTTGGTGGCTT	AGCTCTCATGCCGTAGCATT	60 °C
Contig701.1	Y	CCCTGTTGCACACCTTTCTC	TAACGCAAGAGAGAGCCAAA	60 °C
Contig736.1	N	TTTGAGGGACACAAGACCAAC	TGCAACGTTTTTGAAAGTGG	60 °C
Contig736.2	N	TTTGAGGGACACAAGACCAAC	TGCAACGTTTTTGAAAGTGG	60 °C
Contig744.1	N	CAGGCACATTCACACCAAAC	AGTTCAACGCCTGATGGAGT	60 °C
Contig744.2	Y	ATCCAGTGCTCTGCTTGGTT	AATGTTGGCATGTTTGCTCA	60 °C
Contig748.1	N	AGCAGTGCAAATTTCAGCCT	AGTTCTCCCTGTGCGAGTGT	60 °C
Contig773.7	N	GGAACCGATCCCAGGACTAT	CTATGCAACGGATTTGGGAT	60 °C
Contig786.3	N	AATGTCTCTTTGGGGGCTCT	ACATTTTGGGATGCAAAAGC	60 °C
Contig788.3	N	GCGTCAAGTGATGTGCTGAT	CGGTGTGTTCTGAAATCCCT	60 °C
Contig789.1	N	TCATCCCAAAACACAGCATC	GACACGGAACACGAGACAAG	59 °C
Contig804.3	Y	TTGACTCATCTTGCACTGGG	TGCGGAAGCTTGGATTAACT	60 °C
Contig823.3	N	GGACGTGAAGTTTGGAGAGC	AATGTGTGGGTTTGGGTCAT	60 °C
Contig842.2	N	TCTCCATTGCGAAGAAATCC	ATGCTAGTCATGTCACGGCA	60 °C
Contig847.1	N	CTTCCACAGTGACCAGGGTT	CCAGCCACGATAACCAGTCT	60 °C
Contig851.1	N	CACTCCAGGCTGAAATGGTC	AGAGTGGGAGCAGCTGGATA	60 °C
Contig851.3	N	ACCCCACTTCCATGAGTGAC	GAAACCATCAGGTGTGCTGA	60 °C
Contig852.3	N	AGAAGCGTTGTCCAGCATTC	GGTTGGTGAATGTCGCTCTT	60 °C

Continued on next page

Table A.1 – *Continued from previous page*

Contig name	Female-specific marker?	Left primer	Right primer	Annealing temp
Contig880.3	Y	AGCTTTCTGAAGCGTTGCTC	TTCCTCACTGCGTTCATATTG	60 °C
Contig890.1	N	GCTGCCTTGGTTTCTACAGG	ATGGCAAAAGAGAACCGATG	60 °C
Contig922.5	N	CCCAGTCACTTCCACTCCAT	TTTGCCCGTTTATTTTCGAC	60 °C
Contig925.2	N	GGCCAACTCATGTTACCGTT	AGATTCGTGGAAGATCCGTG	60 °C
Contig974.2	N	GGGAGTGATCAGACCGCTAA	ATAGTGACCCATAGCGCTGC	60 °C
Contig1003.1	N	TGCTATCACTGTGGCACCTC	CATGGAGATCAACATCGTCG	60 °C
Contig1039.3	Y	CAAGTGTGTGCCCTTCCTGT	CTACCTGGCAAACCTAGGCA	60 °C
Contig1086.1	N	GGACAGTGTGGCCATTTTCT	TCCAAAGCAGAAGCCCTAAA	60 °C
Contig1100.2	N	GGCAAGAGGATGGATCAAGA	CGGGAGGCATAGAAATCAAA	60 °C
Contig1103.1	N	CACCAATTCACGCCATACTG	TTTCTTCTCATCTGCACCCC	60 °C
Contig1107.2	N	GAAGAGGTTGCTGCACTCGT	TTTCAACGACAGTGATGGGA	60 °C
Contig1131.1	N	TGGGTCTTGTCTGTTTGACG	TGGCAAGAATACTGTCAAATCC	60 °C
Contig1132.1	N	TGGTGCTGGTTATGGATGAA	TGATCTACGGGTGGTCTTCC	60 °C
Contig1151.6	N	GCTTTGGGTGAGCAAAAGAG	GAAAGCGAGGTGCGTAAAAG	60 °C
Contig1158.3	N	ATATTTTTCCGTGGCACGTC	TCCTATCACTCGCAGGCTTT	60 °C
Contig1180.2	Y	TGAAATCCACCACCCTCAAG	GCCAACCAATAACAGCTTCC	60 °C
Contig1194.1	N	CGGCTGTGCAGACTATGAAA	GTCCCCAGGCAGTAGAACAG	60 °C
Contig1209.1	N	ACAAAGAGAGCGAAATGGAG	ACCTTGAATAGAGCCTGAAATG	57 °C
Contig1220.2	Y	CCGCATAGCTTTTCATCAGTT	AGAGACGAGCCAGAAGATCG	60 °C
Contig1231.2	N	TGTTTCTGTGCAGAACCCAG	TCACATGAAAAACTGCCACC	60 °C
Contig1240.3	N	GCTATGAAACCGATCCCGTA	GTCCTACTGACTTCGCTGGC	60 °C
Contig1249.1	Y	TTCACCCAAATCTTAACAGGG	TGTTGTAACAGGGTGTATCCCA	60 °C
Contig1260.1	N	CTGCTTACGGTGGGTGAAAT	TGTAGCGCAGACACATAGCA	60 °C
Contig1286.2	N	CATAGAGGTGGGAACCCAAA	GTGGAAGCTGTATGTCCGGT	60 °C
Contig1311.3	Y	GCCTTGAACTTGGTGGTTGT	CTACGGCAGCCGAGCTAGTA	60 °C
Contig1314.1	Y	CCCACCACACAGGCAATAG	GATGCAAATTCCGCATGTT	60 °C
Contig1327.4	N	ATTTCTCAGGGGTGCTGATG	AGTGGGAAGACCATTTGTGC	60 °C
Contig1351.5	N	GTGCACATATCCTTCCACCC	GCAGCACTTCGATCATTTCA	60 °C
Contig1405.4	Y	GGGAACAATTGCTACCTGGA	CACAAACCGGTTGGGTAAAG	60 °C
Contig1409.1	Y	GTGGAGGTTAGACAGCCAGC	CCATGCTAGGCTTCCACAAT	60 °C

Continued on next page

Table A.1 – *Continued from previous page*

Contig name	Female-specific marker?	Left primer	Right primer	Annealing temp
Contig1420.1	N	GGGAGACCTGATGCCAATAG	TGGATCACAAATGGAGCAAA	60 °C
Contig1451.1	N	GAGCTGTTCTGTAGGGCTGG	TTCTGCTTCCCATAAATCCG	60 °C
Contig1460.1	N	TGGGTTTCAGGTTCAGGTTC	TCTGGTTGTGTGCTTGTTTG	59 °C
Contig1475.2	N	AAAAAGAGGAACAATGCGTTAC	AAATCGTTATCAGCCTGCTC	60 °C
Contig1486.1	N	ATCTCCAGGACAGGAGGACC	ATAGGGAAGCCCTCACAGGT	60 °C
Contig1506.2	N	AATCTTGGTGGCCTGTTGAC	CGTGACCTCAAGTGGTGCTA	60 °C
Contig1510.1	Y	GGGTCTTGCTGGAAAACTGT	GCGTGGAAAATGCCACTATT	60 °C
Contig1511.1	Y	TTCACAGATCCTCAGGGAGC	CACAGTAAAGCCCAGAGTTGC	60 °C
Contig1515.2	Y	CATGACATCATGCGATAGGC	TAATCGGATTGGCAAACCTC	60 °C
Contig1520.1	N	CGTAAGCCAGATTCACCCAT	TAAAAGCCCTCTTCGTGCTG	60 °C
Contig1542.2	N	CCTCAGACGTCACAACATGC	AAACCCTCTTTTGTTGCCCT	60 °C
Contig1589.2	N	ATTCCCATCTGGTTCAGTCG	AAAAATGGATGCGATTCAGC	60 °C
Contig1592.1	N	CAGCGATTGATGGTCTTGAA	CTAGGGAAGGTGCGAACAAG	60 °C
Contig1634.3	Y	TGAACCTGGCTCCTCTGAAC	CCTCATTTTGAAGGCAAACC	60 °C
Contig1672.1	N	AAATGACAGGCTGTTGGTCC	CTTGCACGTGTGTGTGAATG	60 °C
Contig1690.3	N	CCATTCGTAGAACCCCTCAG	CAGCCCACATAGAGTGCGTA	60 °C
Contig1739.1	Y	CGTCCAGATGCACAAGCATA	GGTTGGACATATACGCCTGG	60 °C
Contig1755.2	N	AATGGGTGTGTGAGAGGAGAG	CTGCTGCCTTTGTGGATTG	60 °C
Contig1763.3	Y	CAAGAACCCAAGCAAGGTGT	GGGGTTGTGTAGGGAAAGGT	60 °C
Contig1778.2	N	ATGTGGGGTTATATGGCTGC	AACCCCACCTTATGGAGGTC	60 °C
Contig1796.2	N	ACTTCACCCTTGCTCACCAT	GCACATGGTGCTATCCAGAA	60 °C
Contig1835.1	Y	GTGGCTTCTTGCTCTCATGC	GGAACAGGGGACAAATCTCA	60 °C
Contig1835.2	Y	CAAGATCTGGGGATGTTGCT	CAATCCCAGGGTTGAAATGT	60 °C
Contig1941.2	N	ACGTGAATTGTGTTCTGTAGCTTG	GCTGGTGGAAATGGTGAATG	61 °C
Contig1948.1	N	CTCCAGGCCAATTTGACAAT	TTATCTGGAAGGAGGCAAGC	60 °C
Contig1964.5	N	CCTGCTTTTGTGACCCCTAA	ACCCTGATGCCATTTTGTCT	60 °C
Contig1974.1	Y	CTCGTCACCATTCCCAGTCT	AGAAATTGAAAAGGAGCGGG	60 °C
Contig2014.1	N	CAGAAATTGCAGAAGCTCCC	GCACCTTGTTAGCCTTGCTC	60 °C
Contig2033.1	N	CCAGTGCTACTCCATGCTCA	AGCAGGGGATTTTCCTGATT	60 °C
Contig2041.1	N	GTAAACTGACCTTCATTCATCTCTC	TTTCTTCCTTTGCCACACTC	58 °C

Continued on next page

Table A.1 – *Continued from previous page*

Contig name	Female-specific marker?	Left primer	Right primer	Annealing temp
Contig2053.1	N	CCAACAACAGGAAGACCAAATC	TCAGCACGGAGTAGAATGACC	60 °C
Contig2178.2	N	AGGGCAAGGGAAGAATCG	ACATTTGACGGCCATTTGAC	60 °C
Contig2208.3	Y	GAGTCCATCGAAAGAGACGC	TGCTGGGACATGGTAAGTCA	60 °C
Contig2256.1	N	ATTTGGGGACAGGCTTTAGG	AATCCTTTGCTTGGCTACGA	60 °C
Contig2319.1	N	AGTCTCTCACCCACCCAATG	GCCACATTTTCCAACCTCTG	60 °C
Contig2325.3	N	ATTAAGGGAACACGCCACAC	GAAGCCGGAATAAAGTGGCT	60 °C
Contig2561.3	N	CTCGAGCAGCTCCACGTAGT	CTACAAATGTTGGGATGCGA	60 °C
Contig2585.1	N	TGTGTGGTGAGACTTCAGCC	CAAATAATTGGGTGTTGCCC	60 °C
Contig2603.3	N	TGCACGGCATCATCTAAGAG	CCCCAAAGGGATGAATTTCT	60 °C
Contig2615.1	N	GCTCTGATTCAAATGGTGCC	GGCAGTTGCAATAACTGGGT	60 °C
Contig2639.1	N	ATTTGAGCTTTGTTGGGGTG	CTGCAGTAAAACCGTGCTTG	60 °C
Contig2650.1	Y	CCTTTAACGACCCTCAACCA	CCCATTAATTACAGCCCCCT	60 °C
Contig2784.2	Y	CGGATTCACATCCAAACAGA	GACTGCTGCTTTAGATGCTTTG	60 °C
Contig2802.2	N	GTCACTGAACCAAACTGGGC	CGTCAAACAGACAAGACCCAT	60 °C
Contig2817.2	N	GCGACCTAACAGTCTGCTCC	GCTCCTTTATCCGGTTTCGT	60 °C
Contig2838.2	N	GCATTAGTATCCCTGCGAGC	TCAGAGGAAGCCAAGATGCT	60 °C
Contig2842.1	N	TAACCACGTGACCCCCTTTA	AGGTGACCCCAATGGAAGAT	60 °C
Contig2870.2	N	TATGGGATAATGGCTCCGTC	CAAAATGCCTCCCACAAAGT	60 °C
Contig2879.2	N	CATCCCATCAAAGAGACCATC	GAACATCGCAAACCAACCTC	60 °C
Contig2883.1	N	TCCAGGGATGCTGTACTTTG	AGCCAAGTCATACCTCCAGC	60 °C
Contig2890.1	N	TGCCCCAAGTCTAAAAATGC	CTATGGTCACCCCTTCTCCA	60 °C
Contig2952.1	N	TGTCAAGGCTTTGGTTCACA	GTGAGCGGTGTGTTAGAGCA	60 °C
Contig2958.1	N	TTTTCACCCCAGAAATGAGC	TAAGCGTACCCAAAACCCAC	60 °C
Contig3021.1	N	GCAGGCTGGGTTAGTCAGAG	GTGTGGCAGATCACAGGAGA	60 °C
Contig3081.1	N	GAAGCAGCTCTGTCGCTTTTA	GGGGTACTGGGAGCCATTAT	60 °C
Contig3118.2	N	AGGACAAAGGCAAAGAGCAG	GAGGAGGAGGAGTGGTTCAG	59 °C
Contig3129.1	N	CATTGTACACCCGTCAGCAC	GGACAGCCTTGTGGAGATGA	60 °C
Contig3141.3	N	AGTGACGCTTCGTGAGAACC	ACATAAGTGCCCCTCAGCC	60 °C
Contig3153.2	N	GCGGGGAAAGATAAAAGAGG	TGAGACACATTAGATGGGGG	60 °C
Contig3189.1	Y	ACCCCAATGACCTCGTGTTA	GGAGGACCGGTATTCCAAAG	60 °C

Continued on next page

Table A.1 – *Continued from previous page*

Contig name	Female-specific marker?	Left primer	Right primer	Annealing temp
Contig3211.1	Y	ATACATGACCGAAGGGACCA	CATGCCAGTGTTGCTGTACC	60 °C
Contig3211.2	N	AGTGCACATACCAGAGGGCT	GTGCCGAAATCTCCATCAGT	60 °C
Contig3211.3	Y	TTTCCGCTCAGTCAGCTTCT	GCATACGCTTGCCCTGTACT	60 °C
Contig3218.1	N	CTCCTCTTTACCTGCGTTGC	ATACTGCATGGGTACAGGGC	60 °C
Contig3262.1	Y	CACATGACGCACCTTTATGG	GGCTGTTCTTCCAGGTTTGA	60 °C
Contig3296.1	Y	GTGGAGGCATCAGTAGGGCT	TCCCAGTGGATGGATAAGGA	60 °C
Contig3298.1	N	TGCCCTGTCTCGCATATACA	CGTCTCGGTTTGGACATTTT	60 °C
Contig3318.2	N	GTGAGAGCTAACAAAGCCCG	GAGCGCTAGAACACCCCATA	60 °C
Contig3346.2	N	TCTCCGTGGCTTTATTTTGG	ATGTTGGTGCAAAGAGACCC	60 °C
Contig3436.4	N	AACAATCCCATAGCCCAGTG	GAGACATTTGGGGGTCTGAA	60 °C
Contig3591.2	N	TATAAGGGGCGTGGTCAAAG	TCCACATCCCCTTTCTATGG	60 °C
Contig3623.1	Y	GTACATGGGACCAACCCTGT	TGTATCAGGCGTGAGCGTTA	60 °C
Contig3641.1	N	TGTTATGAATGGGTAAGGGATG	GGGAGGTAGCAGTCCAGAAG	59 °C
Contig3644.1	Y	TTGCCTTATTTTCTGTGCCC	CCGCCTTCTACCGTGTCTAC	60 °C
Contig3707.2	N	TTTCTCTACCCTTCAACCCA	TTAAAAGCACGTTCACGCAC	60 °C
Contig3734.1	N	CATTGAGCCAACCAGTTGAA	GGTTGGCTGTAAGGACAGGA	60 °C
Contig3835.2	N	AACTTGGATCAGTGGTTGGC	TGAGCTCTGGTTTGGGACTT	60 °C
Contig3951.1	N	TAGGGTGATGGGAGGTGAAA	CCATACAAGCCTCCCCTACA	60 °C
Contig3971.2	Y	GCAGCCCTTCCTGAACATTA	CCAGATGTTAATTCCGTGGG	60 °C
Contig4134.1	N	GATTTCTTCGGAGGCTTGC	TGTAGGCCTTGGGCTTATTG	60 °C
Contig4139.1	N	TTTAAAGCCTGTGTCCACCC	ACAGGTGCGACCCTAATGAC	60 °C
Contig4166.1	N	AGCAGAAGGCGACATTTTGT	GCTGAAGCTGGTACAGGGAA	60 °C
Contig4220.1	N	AATTCTTTACCCACGTGCAA	TTTCCTTGAGGTGTGAGCCT	60 °C
Contig4244.2	N	GCCCCATTCCAAACCTTATT	CATATCTGCCCCAAATCCAC	60 °C
Contig4303.1	N	CTCATAACAGCCCCTCGAAC	TTTTGGGTGTTTTGGGGTAG	60 °C
Contig4350.3	Y	CGGCGAGTGGGTGTATCTAT	CAGCCACCTTCAACATCAAC	60 °C
Contig4351.2	Y	AGGCACCCGTATTCACAAAA	CCCCATAAGGGGCAAACTAT	60 °C
Contig4483.2	N	GTGCTAAGGGGATTTCGTGA	TCACGGGTCTTTAGGGTCAG	60 °C
Contig4605.2	Y	AAAAATCCTTTTGAGGCAGTCA	GCTTTACATTTTGGCTCCCA	60 °C
Contig4688.1	N	TTGTTCCTGAGGGGATTGTC	GCACGGTGAGGGAATAAAAA	60 °C

Continued on next page

Table A.1 – *Continued from previous page*

Contig name	Female-specific marker?	Left primer	Right primer	Annealing temp
Contig4707.1	N	AAATGTCGCGAAAGAAAAGC	CACTGGGTCTGGCGAACTAT	60 °C
Contig4708.1	N	CACGATTCAGAGGTCAAGCA	ATGAGTTTTGCACCCTCCAG	60 °C
Contig4746.1	N	CAAAGGCTTTTGGGTAGCAG	CTGTAGCTGCGTGCATTCAT	60 °C
Contig4748.1	N	CAAAGGCTTTTGGGTAGCAG	CTGTAGCTGCGTGCATTCAT	60 °C
Contig4915.3	N	AATGCAACGCAGACCTAACC	GGAGCTTGGAAGACTGCGTA	60 °C
Contig4937.1	N	ACATGTGCCATACACACGCT	GTCTGCGTGGTAAAGGGGT	60 °C
Contig4944.1	N	ACCTGCACAATTTCCCATCT	GCCACAATGTATGGCAAAGA	60 °C
Contig5029.1	N	GAACGGGCTTAAAGGAAAGG	AGCTGCCCGTGAAAACTAGA	60 °C
Contig5053.2	N	GCGTGTTGTTGTATGTTGGG	ACTCGTGACCGTAGGCAAAT	60 °C
Contig5100.1	N	ACGCTGTTTTGTCCTAACGC	ACTTCAGTTCAAACACCCCG	60 °C
Contig5216.2	N	TATTTGTTGCCATGTGGGTG	ATATGGACCCAGAGCACGTC	60 °C
Contig5266.2	N	GGGAGCCCAGTTTACCTCAT	GAGGCCCAGTAAGAGCTGTG	60 °C
Contig5288.1	N	TTTGTGGCAAGACACGGTAG	CCCGCCCTACAATTATGATG	60 °C
Contig5418.1	N	TCCTGCTGATGTCCCACATA	CTTCCCAACCACAAGAGCAT	60 °C
Contig5696.1	N	CCTGTTCCTGTTCCTGTTCC	GAATCCCTTGTGCCTGAAGA	60 °C
Contig5780.2	N	GCTGGTTGGGTTGTGGTC	AAGTACCGCCCCATATCACA	60 °C
Contig5876.1	N	CTTGGGTTTTGGCTGATGAT	AAAATCCATAAGGGGTTGCC	60 °C
Contig6113.2	N	CTCATCCCGCTCCAAATC	TGCTATTACCCTACAGTGGCTTC	59 °C
Contig6373.1	N	ATACCTGCTCTCCCCTCCAA	AGGGGTTTATGGACCCAGAC	60 °C
Contig6408.2	Y	GACTCTCCTCCACCATCTTCTG	ATTCACCACACTGGGCTTTG	61 °C
Contig6502.1	N	AGGTCACCGACAGCGTAAAC	AACCCACAATCCCATCTGAA	60 °C
Contig6513.2	N	GTTGATAGGGACCTGACGGA	ACACTTGTAGCCGTGGGAAG	60 °C
Contig6764.1	N	GACCGCTCACTCACTGTCCT	TGTCATTGCTTGGGTGATGT	60 °C
Contig6895.1	Y	ATACCAGTTGCAGGGCTACG	CAAGACCTTTCTGCTCCCAT	60 °C
Contig6910.1	N	CCCTTTTACCACGAGTTCCA	GAATGCATTGGGCACAGAG	60 °C
Contig7121.1	N	ACGGGTATGAGGTTCAATGC	GCACCTTAATGCCTCTCCAC	60 °C
Contig7339.1	N	CTGTCTGCAGAATGAGCCTG	CAAGTGGAGATGGGACCTGT	60 °C
Contig7456.2	N	CCCGCACCAATATAAGCATT	GACCCATCATGGTGACAGTG	60 °C
Contig7563.2	N	AGATGTTGGGGGCTCTTAGG	AGTATCGCCCCATTGACAAC	60 °C
Contig7578.1	N	ACTTTGGAGACCCCTCACAC	GGGGTGTTGGGAGGATCTAT	60 °C

Continued on next page

Table A.1 – *Continued from previous page*

Contig name	Female-specific marker?	Left primer	Right primer	Annealing temp
Contig7896.1	N	GGAGGGATTTGGTTTGGTTT	GCCCCACAACCTTCCTCTAT	60 °C
Contig7947.2	N	CGAGGAGGAGGAAGATGATTG	CCACGCCCACAGTAGAAGAG	61 °C
Contig8091.1	N	CCACCTGTGAGATCTTTGGC	TCTTGAAATAGGTGGGGTCG	60 °C
Contig8209.1	N	AAGGGAAGCAGATGGAAGTG	GGCTGAGGAAGGGACAATC	59 °C
Contig8227.2	N	ACTGGCCCCACCAACCTAC	GAAGCCCATGCCCATCAC	62 °C
Contig8239.1	N	TGCTACCTAGCTGCCTCCTC	ATTGACTGCTGATTGCTCCA	60 °C
Contig8478.1	N	GAATCTCAACAAAGGCCATCA	TCAGACGTGTGGCCAGAATA	60 °C
Contig8600.1	N	CTCAGTGGAGAGGGCAAAAG	GACCTAATCAAGCCCAGCAA	60 °C
Contig8693.1	N	CATCATCCCTCCCATCTTG	CTAAGCATTCCCTTTCACAATC	58 °C
Contig8762.1	N	CTGGATGACCTGAAACGTGA	GATGGCATTCAAGGGTGAGT	60 °C
Contig9315.1	Y	AGCAGTTTATGGGGGAACAG	AACTGCATGCATGTATGTCTCC	60 °C
Contig10123.1	N	CACCCAGTTCTCAAGGCTTC	ATGGGGAAAGGTTTCGAAGT	60 °C
Contig10132.1	N	ACACATCGCCTCATCTTTCC	GTTTCTTCCCAGCACCTCAC	60 °C
Contig10152.1	N	AAGAGAGCGTTTGGGTAGGG	GGGTTTGGGTTGGAATCAC	60 °C
Contig10235.1	N	CACCCAGTTCTCAAGGCTTC	ATGGGGAAAGGTTTCGAAGT	60 °C
Contig10549.1	N	GAATGGGCCTTTTCTTGTCA	TTTGAGAGGCATACACAGCG	60 °C
Contig10935.1	N	GGCAAGAGGATGGATCAAGA	CGGGAGGCATAGAAATCAAA	60 °C
Contig11104.1	N	ATTTAAGGCCCAGAATCCCA	GCAACCAAGTGGACGTTTTT	60 °C
Contig11194.1	N	ATGAAACCCTGTGTCCTCCA	TGATGAAATCTCCCTCCCAT	60 °C
Contig11319.1	N	AAATGACCCCAAAAACCACA	CCATTCTCTCCGTCAACCTC	60 °C
Contig11338.1	Y	CCTAAGTTTGTTCCCGCCTT	ACTTTCCACCATTGCACAGA	60 °C
Contig11523.1	N	ATGAGTTGGCCGATGACTGT	GGTATGATCACTCACGCCCT	60 °C
Contig11635.1	N	TCAGTTTCCAAGTCCGTTCC	CCCTGGGCAGATTTGAAGTA	60 °C
Contig11638.1	N	TCCCTGCCATTCATTTATCTTC	AATCAGTCATCCTGCCAACC	60 °C
Contig11691.1	Y	ACGTGCAGAGTAGCTTGCCT	CCCCATTCCCCATAAACTCT	60 °C
Contig11727.2	N	AGCGTTGAGTAATGGTTGGG	GAGGTTGAGGCCACTGAATC	60 °C
Contig11747.1	Y	AGATCAACTTCAGGGTGGCA	CCTCCTTGCAGATAACGTGG	60 °C
Contig11847.1	N	CTGTGTGAAAGGTGGTTCCC	GATCCCAGGACACATAGCGT	60 °C
Contig11950.1	N	AGAATGTGGAAAAACCCGC	GGCTCTTCTTGAAAGGTTCG	60 °C
Contig11957.1	N	AGCAGATGAGTTCCACCAGG	GGCATGAGATATGTGGGGAC	60 °C

Continued on next page

Table A.1 – *Continued from previous page*

Contig name	Female-specific marker?	Left primer	Right primer	Annealing temp
Contig12399.1	N	GACCCACAGATCCCAGTTTG	AAAACGTGGTATGGGGCAC	60 °C
Contig13219.1	N	AAGGGATCACAGCAACCAAC	TCCCTTTTTCAAAGCGTTTCT	60 °C
Contig13772.1	N	GATGCCAGAAGCGTTTGC	GAGAGGGGATGACTTCGACA	60 °C
Contig13865.1	N	GCATCTGGAGACACCTGGAT	GGAAAGCCTGGGTAATGGTT	60 °C
Contig14186.1	N	TCCCTGCAGCACTGTAAATG	ATACTCCCGAGCATGAGGGT	60 °C
Contig14188.1	N	GGAGCATTCATAGCCACTAAAAC	CCACTTCACAGGGACTTCTTC	59 °C
Contig14260.1	N	TGTTGTGCACGTGGATTCTT	CCGAGAAGTCAGCCTAAACG	60 °C
Contig14439.1	N	GAGAGGTGAAGGTGGTGCTG	AGGAGGACGAGGTGGTGTTG	62 °C
Contig14772.1	N	CTGACGGAAATCTATGGGGA	TTGCCATAAGACTACGGGCT	60 °C
Contig15203.1	N	CTCGTTAGCCTAAACCCACG	CCGGGGGATTTTAATAGAGG	60 °C
Contig15501.1	N	GGTCCTTTCTTCTGTTGCTTG	ACACACTTCACCTTGTCTCTAATG	58 °C
Contig15774.1	N	CCCATGGCCTTTTAGAGGAT	GGTATGAATCCATGCCCTGT	60 °C
Contig15996.1	N	GGCCCTGGTAGTACTCCTCC	AGTGGTGATGGATGCACAAA	60 °C
Contig16112.1	N	ATGGAGGGGGTTGAGGACT	CTGGCAATTGCTACGGAACT	60 °C
Contig16657.1	Y	CCGAGGCAGAAAATAGTGGA	AGCCTTGAGAATCGTGCAAT	60 °C
Contig16713.1	N	GCTGTGGTGGGAAGTGATG	TTGCCTCTTTGACCTCTTGG	60 °C
Contig16781.1	N	AAGCGGGGTGGTGTGTATC	CAAGCAACGGGATGGAAG	60 °C
Contig16914.1	N	ACAGGGTAAGCAGCCTCACA	CATGGATCCTTGTGCTCGTA	60 °C
Contig17240.1	Y	TGTGTCCGAAGAAGATGCTG	GGGCAAAGAACCAATGAGG	60 °C
Contig17520.1	N	TTCCCGCCCGTATTTGTA	GGCCAGGGAGTTAACACTGA	60 °C
Contig17660.1	N	GCATTGATTGGTCTTGCTCA	GGGATGGAAGGGAATCCTAC	60 °C
Contig18173.1	N	TTCTGCACAGACCCACTCAC	GCACATCGTATCCGATCACA	60 °C
Contig18661.1	N	TGCTGCACCGTCAGATTTTA	CAGTCCATCTGCGTGCTTTA	60 °C
Contig18681.1	N	TTTAAAGTGTGTTGCGCTGG	GTAAAGATAAGCCCTCCCCC	60 °C
Contig19045.1	N	TCTAATGGATGGGGTAAGTGTG	TGGTGTTCGGTTGAAATCTG	59 °C
Contig19184.1	Y	TCTTGGACCCTTAATTGCTTG	TGCCATGCTTATGAGTCAGC	60 °C
Contig19387.1	N	TTTGGGGTTTTAGGGGTCTT	CCCCAAAATGACCGAAAATA	60 °C
Contig19481.1	N	CTCCAGGCCAATTTGACAAT	TTATCTGGAAGGAGGCAAGC	60 °C
Contig19668.1	N	AGACGAATGTGAAACGGAGC	AAACTGCACTCCGCTGAAAT	60 °C
Contig19853.1	N	TCAGTTCACGTGTCCTTTGG	GCCAGTAATGAGCTGCCTCT	60 °C

Continued on next page

Table A.1 – *Continued from previous page*

Contig name	**Female-specific marker?**	**Left primer**	**Right primer**	**Annealing temp**
Contig20120.1	N	GCCTTCTGTGCACTTTATGG	GATTGGCCGCTACCTCAGA	60 °C
Contig20335.1	N	CCATCTATGGGCAATCCAAC	ACTCGCTGGTGACTCAGCTC	60 °C
Contig20734.1	N	AGAGCTGTTTCCAATCTGCG	CCTGTTCAATTAACGGCTGG	60 °C
Contig21126.1	N	AGAGCCCTTACAGGGTCCAT	ACCCCATAAAGTCCCCTGAC	60 °C
Contig21340.1	N	CCCCAAGGGACCATTTATTT	AGGAGGGGCAACTTTTTCTT	60 °C
Contig21666.1	N	GGAAAGGGACAGGGTAGGAC	TGTGGCTCTACAAAACCCCT	60 °C
Contig21669.1	N	TACGCTCCCTGCTATCTTCC	GCCCTGCTGTTCCTTATTG	59 °C
Contig21870.1	N	CCTGTGTGCTGATTGTATCCTG	CTGTGGCTCGGACTGTTCTC	61 °C
Contig21949.1	N	ACTTGAGCAGCTGTCGTGGT	ACTCCAATCTCGACCCACAC	60 °C
Contig21953.1	N	GGAGCGGTTGGTTTGTCTG	CGCCTTGGGTTTGTTTCTG	62 °C
Contig22089.1	N	ATTTGAGTTTGGCTGCAAGG	GCCCATGTGACTAAACAGCA	60 °C
Contig22504.1	N	TTATTATCCCCACCGGGAAG	CGATTAAAAGTGCAGGGAGC	60 °C
Contig23065.1	N	TTATGGAAATTGGAGCCAGG	TTTGAAAGCATTCTGCACCT	60 °C
Contig23455.1	N	ATTAAGGGAACACGCCACAC	GAAGCCGGAATAAAGTGGCT	60 °C
Contig23497.1	N	GCAGATAGGAGGTGGAGGAG	TCAGAGAGGTGAGGTTGTGG	59 °C
Contig23624.1	N	AGAAGGTTGGGTGAGTGCTG	CATCTGGGATTTGTTGAATGG	60 °C
Contig23699.1	N	GTGCCTCATTGCACAAACAC	TAATGCTTGTGCATCTTCCG	60 °C

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

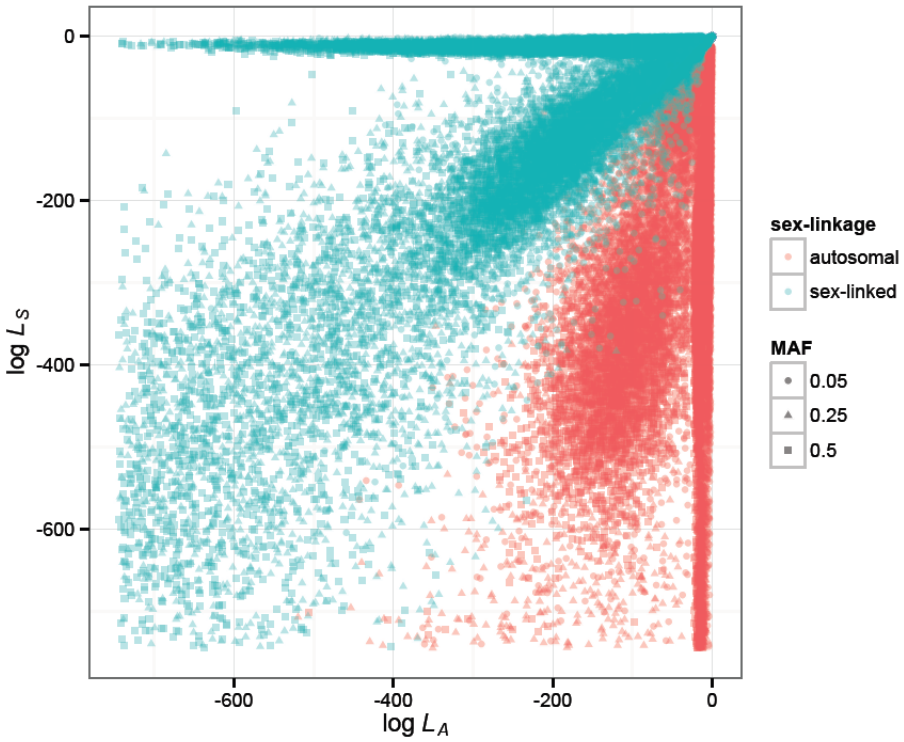

Figure B.1: Confirmation that pedigree likelihoods can be used to identify sex-linked sites. We simulated autosomal (pink) and sex-linked (blue) SNPs with medium to high quality genotypes and 0-20% missing data. For each SNP, we plot the likelihood of the pedigree under an autosomal model of inheritance (L_A) and the likelihood of the pedigree under a sex-linked model of inheritance (L_S). The shape of the points indicates the MAF of the simulated SNPs. Autosomal SNPs and sex-linked SNPs have different pedigree likelihoods. Therefore we can classify SNPs as autosomal or sex-linked based on L_A and L_S.

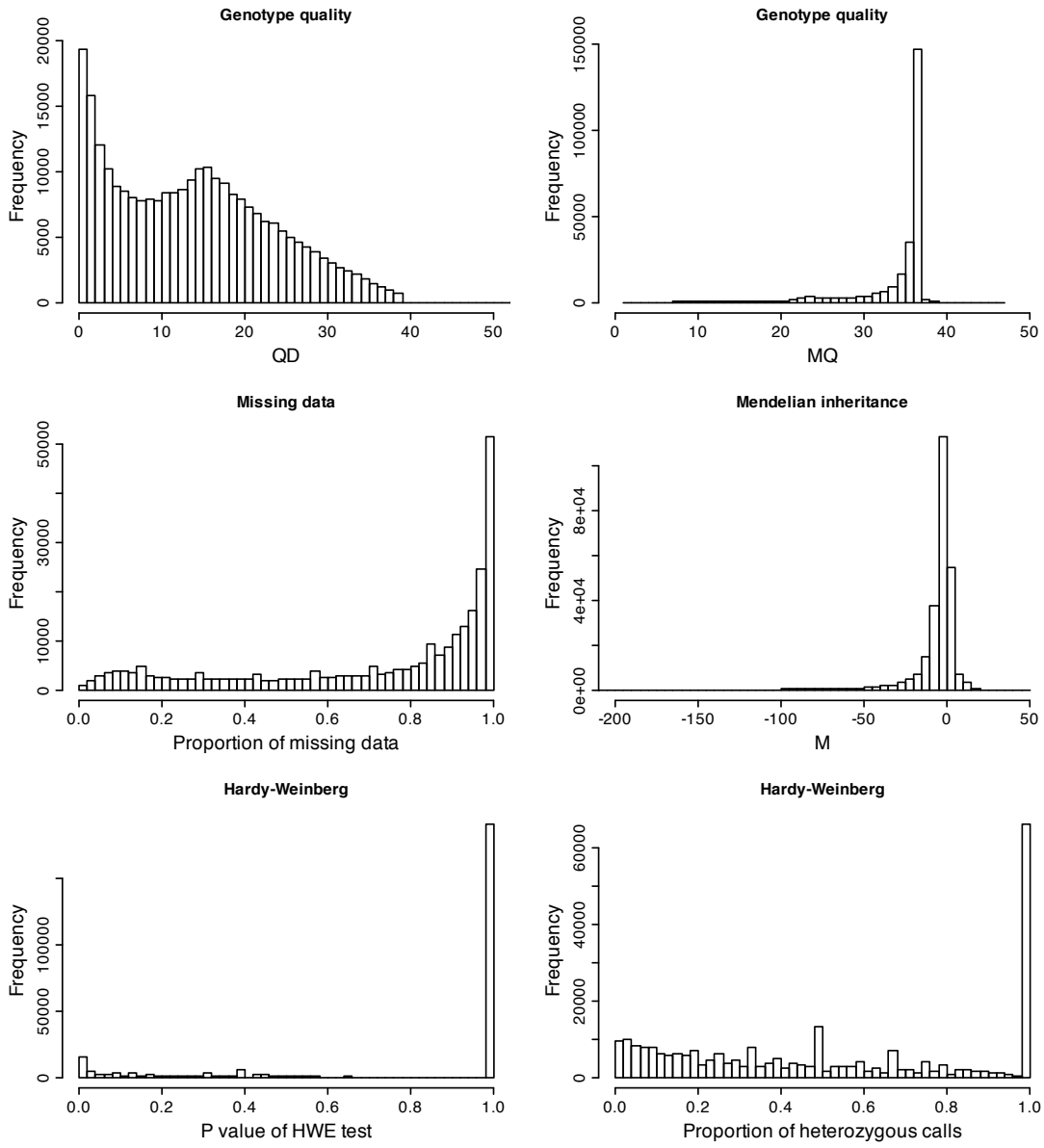

Figure B.2: Distributions of various quality metrics (genotype quality, missing data, Mendelian inheritance, and Hardy-Weinberg) for unfiltered SNPs discovered using GBS in Florida Scrub-Jays.

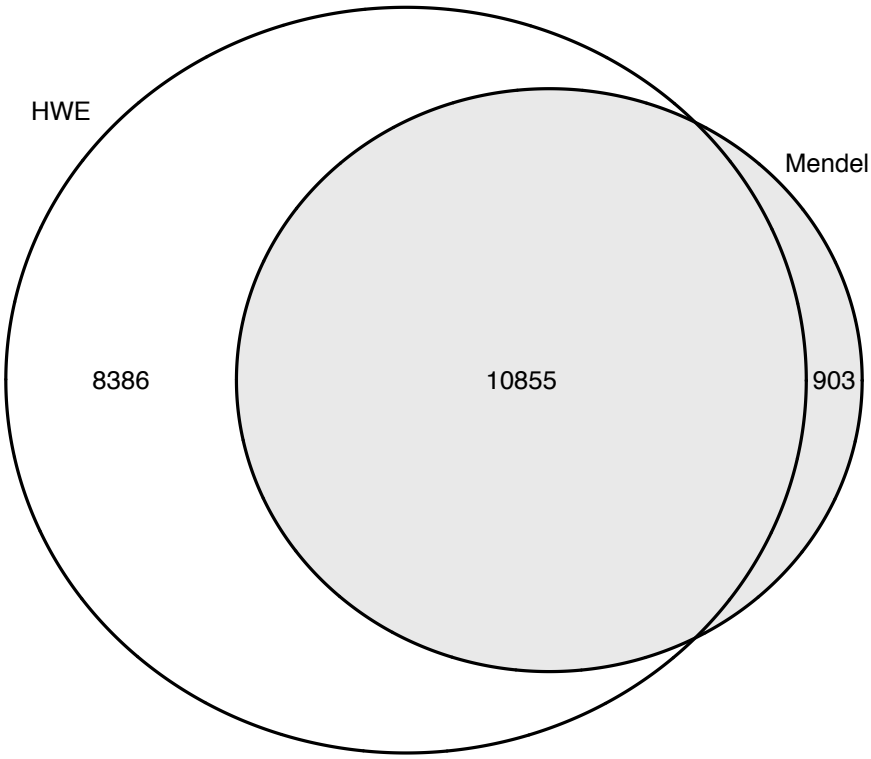

Figure B.3: Number of high-quality SNPs from the real data that pass a Hardy-Weinberg test or the Mendelian inheritance filter. SNPs have already been filtered for quality and proportion of missing data. The Mendelian inheritance filter is more rigorous; 44% of the SNPs that pass the HWE test fail MendelChecker but only 8% of the SNPs that pass MendelChecker fail HWE.

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

Table C.1: Number of SNPs after each filtering step during the Beadchip design process.

Filter	Number of remaining SNPs
thinned to 1 SNP per 100 bp window	41,853
QD > 2	40,310
MQ > 35	37,835
MISS < 0.92	27,956
Mendelian score > -20	26,750
MAF > 0.02	22,575
SNP < 50 bp from start	22,569
excess heterozygosity (HWE test)	22,555
remove triallelic SNP or degenerate nt	22,230
Illumina assay score > 0.7	19,209
remove repeats	19,087
Illumina assay score > 0.781	17,856
MAF > 0.0223	17,629

BIBLIOGRAPHY

Abecasis G. R., Cherny S. S., Cookson W. O., and Cardon L. R. 2002. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics *30*(1)**:** 97–101.

Adams J. R., Vucetich L. M., Hedrick P. W., Peterson R. O., and Vucetich J. A. 2011. Genomic sweep and potential genetic rescue during limiting environmental conditions in an isolated wolf population. Proceedings of the Royal Society B *278*(1723)**:** 3336–44.

Albers C. A., Stankovich J., Thomson R., Bahlo M., and Kappen H. J. 2008. Multipoint approximations of identity-by-descent probabilities for accurate linkage analysis of distantly related individuals. American Journal of Human Genetics *82*(3)**:** 607–22.

Alkan C., Coe B. P., and Eichler E. E. 2011. Genome structural variation discovery and genotyping. Nature Reviews Genetics *12*(5)**:** 363–76.

Alkan C., Sajjadian S., and Eichler E. E. 2011. Limitations of next-generation genome sequence assembly. Nature Methods *8*(1)**:** 61–5.

Allendorf F. W., Hohenlohe P. A., and Luikart G. 2010. Genomics and the future of conservation genetics. Nature Reviews Genetics *11*(10)**:** 697–709.

Almasy L. and Blangero J. 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. American Journal of Human Genetics *62*(5)**:** 1198–211.

Andolfatto P., Davison D., Erezyilmaz D., Hu T. T., Mast J., *et al.* 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Research *21*(4)**:** 610–7.

Andrés A. M., Dennis M. Y., Kretzschmar W. W., Cannons J. L., Lee-Lin S.-Q., *et al.* 2010. Balancing selection maintains a form of ERAP2 that undergoes

nonsense-mediated decay and affects antigen presentation. PLoS Genetics *6*(10)**:** e1001157.

Angeloni F., Wagemaker N., Vergeer P., and Ouborg J. 2012. Genomic toolboxes for conservation biologists. Evolutionary Applications *5*(2)**:** 130–143.

Arnold B., Corbett-Detig R. B., Hartl D., and Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Molecular Ecology *22*(11)**:** 3179–90.

Axelsson E., Smith N. G. C., Sundström H., Berlin S., and Ellegren H. 2004. Male-biased mutation rate and divergence in autosomal, z-linked and w-linked introns of chicken and Turkey. Molecular Biology and Evolution *21*(8)**:** 1538–47.

Ayroles J. F., Hughes K. A., Rowe K. C., Reedy M. M., Rodriguez-Zas S. L., *et al.* 2009. A genomewide assessment of inbreeding depression: gene number, function, and mode of action. Conservation Biology *23*(4)**:** 920–30.

Barton N. H. and Etheridge A. M. 2011. The relation between reproductive value and genetic contribution. Genetics *188*(4)**:** 953–73.

Bates D., Maechler M., Bolker B., and Walker S. 2004. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-6. http://CRAN.R-project.org/package=lme4.

Baxter S. W., Davey J. W., Johnston J. S., Shelton A. M., Heckel D. G., *et al.* 2011. Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PLoS One *6*(4)**:** e19315.

Bensch S., Hasselquist D., and von Schantz T. 1994. Genetic similarity between parents predicts hatching failure: non-incestuous inbreeding in the Great Reed Warbler? Evolution *48*(2)**:** 317–326.

Blouin M. S. 2003. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends in Ecology & Evolution *18*(10)**:** 503 – 511.

Boughton R. K. and Bowman R. 2011. State wide assessment of Florida Scrub-Jays on managed areas: A comparison of current populations to the results of the 1992-1993 survey. Technical report, submitted to the USFWS.

Browning B. L. and Browning S. R. 2011. A fast, powerful method for detecting identity by descent. American Journal of Human Genetics *88*(2)**:** 173–82.

Butchart S. H. M., Walpole M., Collen B., van Strien A., Scharlemann J. P. W., *et al.* 2010. Global biodiversity: indicators of recent declines. Science *328*(5982)**:** 1164–8.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., *et al.* 2009. BLAST+: architecture and applications. BMC Bioinformatics **10:** 421.

Carvalho A. B. and Clark A. G. 2013. Efficient identification of Y chromosome sequences in the human and Drosophila genomes. Genome Research *23*(11)**:** 1894–907.

Carvalho A. B., Vibranovski M. D., Carlson J. W., Celniker S. E., Hoskins R. A., *et al.* 2003. Y chromosome and other heterochromatic sequences of the *Drosophila melanogaster* genome: how far can we go? Genetica *117*(2-3)**:** 227–37.

Casey A. E., Jones K. L., Sandercock B. K., and Wisely S. M. 2009. Heteroduplex molecules cause sexing errors in a standard molecular protocol for avian sexing. Molecular Ecology Resources *9*(1)**:** 61–5.

Catchen J. M., Amores A., Hohenlohe P., Cresko W., and Postlethwait J. H. 2011. *Stacks*: building and genotyping Loci *de novo* from short-read sequences. G3 *1*(3)**:** 171–82.

Charlesworth B., Coyne J. A., and Barton N. H. 1987. The relative rates of evolution of sex chromosomes and autosomes. The American Naturalist *130*(1)**:** 113–146.

Charlesworth D. and Willis J. H. 2009. The genetics of inbreeding depression. Nature Reviews Genetics *10*(11)**:** 783–96.

Chen N., Van Hout C., Gottipati S., and Clark A. G. 2014. Using Mendelian inheritance to improve high throughput SNP discovery. In review.

Chen W., Li B., Zeng Z., Sanna S., Sidore C., *et al.* 2013. Genotype calling and haplotyping in parent-offspring trios. Genome Research *23*(1)**:** 142–51.

Chong Z., Ruan J., and Wu C.-I. 2012. Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads. Bioinformatics *28*(21)**:** 2732–7.

Christiansen F. B. and Frydenberg O. 1973. Selection component analysis of natural polymorphisms using population samples including mother-offspring combinations. Theoretical Population Biology *4*(4)**:** 425–45.

Christiansen F. B. and Prout T. 2000. *Evolutionary Genetics From Molecules to Morphology*, Chapter Aspects of fitness, pp. 146–156. New York, NY, USA: Cambridge University Press.

Clutton-Brock T. and Sheldon B. C. 2010. Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology. Trends in Ecology & Evolution *25*(10)**:** 562–73.

Coulon A., Fitzpatrick J. W., Bowman R., and Lovette I. J. 2010. Effects of habitat fragmentation on effective dispersal of Florida scrub-jays. Conservation Biology *24*(4)**:** 1080–8.

Coulon A., Fitzpatrick J. W., Bowman R., Stith B. M., Makarewich C. A., *et al.* 2008. Congruent population structure inferred from dispersal behaviour

and intensive genetic surveys of the threatened Florida scrub-jay (*Aphelocoma coerulescens*). Molecular Ecology *17*(7)**:** 1685–701.

Cox D. and Oakes D. 1984. *Analysis of survival data*. New York, NY, USA: Chapman & Hall.

Danecek P., Auton A., Abecasis G., Albers C. A., Banks E., *et al.* 2011. The variant call format and VCFtools. Bioinformatics *27*(15)**:** 2156–8.

Davey J. W., Cezard T., Fuentes-Utrilla P., Eland C., Gharbi K., *et al.* 2013. Special features of RAD Sequencing data: implications for genotyping. Molecular Ecology *22*(11)**:** 3151–64.

Davey J. W., Hohenlohe P. A., Etter P. D., Boone J. Q., Catchen J. M., *et al.* 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics *12*(7)**:** 499–510.

Dawson D. A., Darby S., Hunter F. M., Krupa A. P., Jones I. L., *et al.* 2001. A critique of avian CHD-based molecular sexing protocols illustrated by a Z-chromosome polymorphism detected in auklets. Molecular Ecology Notes *1*(3)**:** 201–204.

DePristo M. A., Banks E., Poplin R., Garimella K. V., Maguire J. R., *et al.* 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics *43*(5)**:** 491–8.

Douglas J. A., Skol A. D., and Boehnke M. 2002. Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. American Journal of Human Genetics *70*(2)**:** 487–95.

Eaton D. A. R. 2014. PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. Bioinformatics *30*(13)**:** 1844–1849.

Ekblom R. and Galindo J. 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity *107*(1)**:** 1–15.

Ellegren H. 1996. First gene on the avian W chromosome (CHD) provides a tag for universal sexing of non-ratite birds. Proceedings of the Royal Society B *263*(1377)**:** 1635–41.

Ellegren H. 2000. Evolution of the avian sex chromosomes and their role in sex determination. Trends in Ecology & Evolution *15*(5)**:** 188–192.

Ellegren H. 2007. Molecular evolutionary genomics of birds. Cytogenetic and Genome Research *117*(1-4)**:** 120–30.

Ellegren H. 2010. Evolutionary stasis: the stable chromosomes of birds. Trends in Ecology & Evolution *25*(5)**:** 283–91.

Ellegren H. and Sheldon B. C. 1997. New tools for sex identification and the study of sex allocation in birds. Trends in Ecology & Evolution *12*(7)**:** 255–9.

Ellegren H. and Sheldon B. C. 2008. Genetic basis of fitness differences in natural populations. Nature *452*(7184)**:** 169–75.

Elshire R. J., Glaubitz J. C., Sun Q., Poland J. A., Kawamoto K., *et al.* 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One *6*(5)**:** e19379.

Elston R. C. and Stewart J. 1971. A general model for the genetic analysis of pedigree data. Human Heredity *21*(6)**:** 523–42.

Epstein M. P., Duren W. L., and Boehnke M. 2000. Improved inference of relationship for pairs of individuals. American Journal of Human Genetics *67*(5)**:** 1219–31.

Fisher R. A. 1931. The evolution of dominance. Biological Reviews *6*(4)**:** 345–368.

Fitzpatrick J. W., Pranty B., and Stith B. M. 1994. Florida Scrub-Jay statewide map, 1992-1993. Archbold Biological Station, Lake Placid, FL, USA.

Foerster K., Coulson T., Sheldon B. C., Pemberton J. M., Clutton-Brock T. H., *et al.* 2007. Sexually antagonistic genetic variation for fitness in red deer. Nature *447*(7148)**:** 1107–10.

Foote S., Vollrath D., Hilton A., and Page D. C. 1992. The human Y chromosome: overlapping DNA clones spanning the euchromatic region. Science *258*(5079)**:** 60–6.

Fox G. A., Kendall B. E., Fitzpatrick J. W., and Woolfenden G. E. 2006. Consequences of heterogeneity in survival probability in a population of Florida Scrub-Jays. Journal of Animal Ecology *75*(4)**:** 921–927.

Frankham R. 1996. Relationship of genetic variation to population size in wildlife. Conservation Biology *10*(6)**:** 1500–1508.

Frankham R. 2005. Genetics and extinction. Biological Conservation *126*(2)**:** 131 – 140.

Frankham R., Ballou J., and Briscoe D. 2003. *Introduction to conservation genetics*. Cambridge, UK: Cambridge University Press.

Fridolfsson A.-K., Cheng H., Copeland N. G., Jenkins N. A., Liu H. C., *et al.* 1998. Evolution of the avian sex chromosomes from an ancestral pair of autosomes. Proceedings of the National Academy of Sciences of the United States of America *95*(14)**:** 8147–52.

Fridolfsson A.-K. and Ellegren H. 1999. A simple and universal method for molecular sexing of non-ratite birds. Journal of Avian Biology *30*(1)**:** 116–121.

Gagnaire P.-A., Normandeau E., Pavey S. A., and Bernatchez L. 2013. Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (*Coregonus clupeaformis*). Molecular Ecology *22*(11)**:** 3036–48.

Garvin M. C., Tarvin K. A., Stark L. M., Woolfenden G. E., Fitzpatrick J. W., *et al.* 2004. Arboviral infection in two species of wild jays (Aves: Corvidae): evidence for population impacts. Journal of Medical Entomology *41*(2)**:** 215–25.

Gautier M., Gharbi K., Cezard T., Foucaud J., Kerdelhué C., *et al.* 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Molecular Ecology *22*(11)**:** 3165–78.

George A. W., Visscher P. M., and Haley C. S. 2000. Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. Genetics *156*(4)**:** 2081–92.

Gordon D., Leal S. M., Heath S. C., and Ott J. 2000. An analytic solution to single nucleotide polymorphism error-detection rates in nuclear families: implications for study design. Pacific Symposium on Biocomputing**:** 663–74.

Grabherr M. G., Haas B. J., Yassour M., Levin J. Z., Thompson D. A., *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology *29*(7)**:** 644–52.

Green E. D. 2001. Strategies for the systematic sequencing of complex genomes. Nature Reviews Genetics *2*(8)**:** 573–83.

Griffiths R., Double M. C., Orr K., and Dawson R. J. 1998. A DNA test to sex most birds. Molecular Ecology *7*(8)**:** 1071–5.

Grossman S. R., Shlyakhter I., Shylakhter I., Karlsson E. K., Byrne E. H., *et al.* 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science *327*(5967)**:** 883–6.

Haldane J. and Jayakar S. 1963. Polymorphism due to selection of varying direction. Journal of Genetics *58*(2)**:** 237–242.

Hansson B. 2004. Marker-based relatedness predicts egg-hatching failure in great reed warblers. Conservation Genetics **5:** 339–348.

Hartl D. and Clark A. G. 2007. *Principles of Population Genetics*. Sunderland, MA, USA: Sinauer Associates.

Heath S. C. 1997. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. American Journal of Human Genetics *61*(3)**:** 748–60.

Hedrick P. W. 2001. Conservation genetics: where are we now? Trends in Ecology & Evolution *16*(11)**:** 629–636.

Hemmings N. L., Slate J., and Birkhead T. R. 2012. Inbreeding causes early death in a passerine bird. Nature Communications **3:** 863.

Henderson R. 1984. *Application of Linear Models in Animal Breeding*. Ontario, Canada: University of Guelph.

Henle K., Lindenmayer D., Margules C., Saunders D., and Wissel C. 2004. Species survival in fragmented landscapes: where are we now? Biodiversity and Conservation **13:** 1–8.

Hogg J. T., Forbes S. H., Steele B. M., and Luikart G. 2006. Genetic rescue of an insular population of large mammals. Proceedings of the Royal Society B *273*(1593)**:** 1491–9.

Hori T., Asakawa S., Itoh Y., Shimizu N., and Mizuno S. 2000. *Wpkci*, encoding an altered form of *PKCI*, is conserved widely on the avian W chromosome and expressed in early female embryos: implication of its role in female sex determination. Molecular Biology of the Cell *11*(10)**:** 3645–60.

Hoskins R. A., Carlson J. W., Kennedy C., Acevedo D., Evans-Holm M., *et al.* 2007. Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. Science *316*(5831)**:** 1625–8.

Hoskins R. A., Smith C. D., Carlson J. W., Carvalho A. B., Halpern A., *et al.* 2002. Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. Genome Biology *3*(12)**:** RESEARCH0085.

Hudson M. E. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. Molecular Ecology Resources *8*(1)**:** 3–17.

Hughes J. F., Skaletsky H., Pyntikova T., Graves T. A., van Daalen S. K. M., *et al.* 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature *463*(7280)**:** 536–9.

Ingvarsson P. K. 2001. Restoration of genetic variation lost - the genetic rescue hypothesis. Trends in Ecology & Evolution *16*(2)**:** 62–63.

International Chicken Genome Sequencing Consortium 2004a. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature *432*(7018)**:** 717–22.

International Chicken Genome Sequencing Consortium 2004b. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature *432*(7018)**:** 695–716.

Itoh Y., Hori T., Saitoh H., and Mizuno S. 2001. Chicken *spindling* genes on W and Z chromosomes: transcriptional expression of both genes and dynamic behavior of spindlin in interphase and mitotic cells. Chromosome Research *9*(4)**:** 283–299.

Itoh Y. and Mizuno S. 2002. Molecular and cytological characterization of *Ssp*I-family repetitive sequence on the chicken W chromosome. Chromosome Research *10*(6)**:** 499–511.

Jones A. G. and Ardren W. R. 2003. Methods of parentage analysis in natural populations. Molecular Ecology *12*(10)**:** 2511–23.

Kaeuffer R., Coltman D. W., Chapuis J.-L., Pontier D., and Réale D. 2007. Unexpected heterozygosity in an island mouflon population founded by

a single pair of individuals. Proceedings of the Royal Society B *274*(1609)**:** 527–33.

Kahn N. W., St. John J., and Quinn T. W. 1998. Chromosome-specific intron size differences in the avian CHD gene provide an efficient method for sex identification in birds. Auk *115*(4)**:** 1074–1078.

Kang H. M., Sul J. H., Service S. K., Zaitlen N. A., Kong S.-Y., *et al.* 2010. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics *42*(4)**:** 348–54.

Keller L. F., Jeffery K. J., Arcese P., Beaumont M. A., Hochachka W. M., *et al.* 2001. Immigration and the ephemerality of a natural population bottleneck: evidence from molecular markers. Proceedings of the Royal Society B *268*(1474)**:** 1387–94.

Kempenaers B., Adriaensen F., Noordwijk A. J. V., and Dhondt A. A. 1996. Genetic similarity, inbreeding and hatching failure in Blue Tits: are unhatched eggs infertile? Proceedings of the Royal Society B *263*(1367)**:** 179–185.

Kinsella J. M. 1974. Helminth fauna of the Florida Scrub Jay: host and ecological relationships. Proceedings of the Helminthological Society of Washington **41:** 127–130.

Kohn M. H., Murphy W. J., Ostrander E. A., and Wayne R. K. 2006. Genomics and conservation genetics. Trends in Ecology & Evolution *21*(11)**:** 629–37.

Kristensen T. N., Pedersen K. S., Vermeulen C. J., and Loeschcke V. 2010. Research on inbreeding in the 'omic' era. Trends in Ecology & Evolution *25*(1)**:** 44–52.

Kristensen T. N., Sørensen P., Pedersen K. S., Kruhøffer M., and Loeschcke V. 2006. Inbreeding by environmental interactions affect gene expression in Drosophila melanogaster. Genetics *173*(3)**:** 1329–36.

Kruuk L. E. B. 2004. Estimating genetic parameters in natural populations using the "animal model". Philosophical transactions of the Royal Society of London. Series B, Biological sciences *359*(1446)**:** 873–90.

Kruuk L. E. B., Clutton-Brock T. H., Slate J., Pemberton J. M., Brotherstone S., *et al.* 2000. Heritability of fitness in a wild mammal population. Proceedings of the National Academy of Sciences of the United States of America *97*(2)**:** 698–703.

Kruuk L. E. B. and Hill W. G. 2008. Introduction. Evolutionary dynamics of wild populations: the use of long-term pedigree data. Proceedings of the Royal Society B *275*(1635)**:** 593–6.

Kurtz S., Phillippy A., Delcher A. L., Smoot M., Shumway M., *et al.* 2004. Versatile and open software for comparing large genomes. Genome Biology *5*(2)**:** R12.

Lande R. 1988. Genetics and demography in biological conservation. Science *241*(4872)**:** 1455–60.

Lander E. S. and Botstein D. 1987. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science *236*(4808)**:** 1567–70.

Leimu R., Mutikainen P., Koricheva J., and Fischer M. 2006. How general are positive relationships between plant population size, fitness and genetic variation? Journal of Ecology *94*(5)**:** 942–952.

Li H. and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*(14)**:** 1754–60.

Li H., Ruan J., and Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research *18*(11)**:** 1851–8.

Liu F., Kirichenko A., Axenovich T. I., van Duijn C. M., and Aulchenko Y. S. 2008. An approach for cutting large and complex pedigrees for linkage analysis. European Journal of Human Genetics *16*(7)**:** 854–60.

Lu F., Lipka A. E., Glaubitz J., Elshire R., Cherney J. H., *et al.* 2013. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genetics *9*(1)**:** e1003215.

Lynch M., Conery J., and Burger R. 1995. Mutation Accumulation and the Extinction of Small Populations. The American Naturalist *146*(4)**:** 489–518.

MacCluer J. W., VandeBerg J. L., Read B., and Ryder O. A. 1986. Pedigree analysis by computer simulation. Zoo Biology *5*(2)**:** 147–160.

Mank J. E. and Ellegren H. 2007. Parallel divergence and degradation of the avian W sex chromosome. Trends in Ecology & Evolution *22*(8)**:** 389–91.

Mardis E. R. 2008. The impact of next-generation sequencing technology on genetics. Trends in Genetics *24*(3)**:** 133–41.

Medvedev P., Stanciu M., and Brudno M. 2009. Computational methods for discovering structural variation with next-generation sequencing. Nature Methods *6*(11 Suppl)**:** S13–20.

Miller M. R., Brunelli J. P., Wheeler P. A., Liu S., Rexroad, 3rd C. E., *et al.* 2012. A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. Molecular Ecology *21*(2)**:** 237–49.

Miller P. M., Gavrilets S., and Rice W. R. 2006. Sexual conflict via maternal-effect genes in ZW species. Science *312*(5770)**:** 73.

Nachman M. W. 2002. Variation in recombination rate across the genome: evidence and implications. Current Opinion in Genetics & Development *12*(6)**:** 657–63.

Nadeau J. H. and Baccus R. 1981. Selection Components of Four Allozymes in Natural Populations of *Peromyscus maniculatus*. Evolution *35*(1)**:** 11–20.

Nadeau J. H., Dietz K., and Tamarin R. H. 1981. Gametic selection and the selection component analysis. Genetics Research *37*(03)**:** 275–284.

Nam K. and Ellegren H. 2008. The Chicken (*Gallus gallus*) Z Chromosome Contains at Least Three Nonlinear Evolutionary Strata. Genetics *180*(2)**:** 1131–1136.

Narum S. R., Buerkle C. A., Davey J. W., Miller M. R., and Hohenlohe P. A. 2013. Genotyping-by-sequencing in ecological and conservation genomics. Molecular Ecology *22*(11)**:** 2841–7.

Nielsen R., Paul J. S., Albrechtsen A., and Song Y. S. 2011. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics *12*(6)**:** 443–51.

O'Connell J. R. and Weeks D. E. 1998. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. American Journal of Human Genetics *63*(1)**:** 259–66.

Ogden R., Gharbi K., Mugue N., Martinsohn J., Senn H., *et al.* 2013. Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. Molecular Ecology *22*(11)**:** 3112–23.

O'Neill M., Binder M., Smith C., Andrews J., Reed K., *et al.* 2000. *ASW*: a gene with conserved avian W-linkage and female specific expression in chick embryonic gonad. Development Genes and Evolution *210*(5)**:** 243–9.

Orr H. A. 2009. Fitness and its role in evolutionary genetics. Nature Reviews Genetics *10*(8)**:** 531–9.

Ott J. 1974. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. American Journal of Human Genetics *26*(5)**:** 588–97.

Ouborg N. J., Angeloni F., and Vergeer P. 2010. An essay on the necessity and feasibility of conservation genomics. Conservation Genetics *11*(2)**:** 643–

653.

Ouborg N. J., Pertoldi C., Loeschcke V., Bijlsma R. K., and Hedrick P. W. 2010. Conservation genetics in transition to conservation genomics. Trends in Genetics *26*(4)**:** 177–87.

Ouborg N. J., Vergeer P., and Mix C. 2006. The rough edges of the conservation genetics paradigm for plants. Journal of Ecology *94*(6)**:** 1233–1248.

Paige K. N. 2010. The functional genomics of inbreeding depression: a new approach to an old problem. BioScience *60*(4)**:** 267–277.

Parchman T. L., Gompert Z., Mudge J., Schilkey F. D., Benkman C. W., *et al.* 2012. Genome-wide association genetics of an adaptive trait in lodgepole pine. Molecular Ecology *21*(12)**:** 2991–3005.

Patten M. M. and Haig D. 2009. Maintenance or loss of genetic variation under sexual and parental antagonism at a sex-linked locus. Evolution *63*(11)**:** 2888–95.

Patterson N., Price A. L., and Reich D. 2006. Population structure and eigenanalysis. PLoS Genetics *2*(12)**:** e190.

Pedersen K. S., Kristensen T. N., and Loeschcke V. 2005. Effects of inbreeding and rate of inbreeding in *Drosophila melanogaster*- Hsp70 expression and fitness. Journal of Evolutionary Biology *18*(4)**:** 756–62.

Pemberton J. M. 2008. Wild pedigrees: the way forward. Proceedings of the Royal Society B *275*(1635)**:** 613–21.

Peterson B. K., Weber J. N., Kay E. H., Fisher H. S., and Hoekstra H. E. 2012. Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLoS One *7*(5)**:** e37135.

Presgraves D. C. 2008. Sex chromosomes and speciation in *Drosophila*. Trends in Genetics *24*(7)**:** 336–43.

Primmer C. R. 2009. From conservation genetics to conservation genomics. Annals of the New York Academy of Sciences **1162:** 357–68.

Pritchard J. K., Stephens M., and Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics *155*(2)**:** 945–59.

Prout T. 1965. The estimation of fitnesses from genotypic frequencies. Evolution *19*(4)**:** 546–551.

Prout T. 1969. The estimation of fitnesses from population data. Genetics *63*(4)**:** 949–67.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A. R., *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics *81*(3)**:** 559–75.

Pusey A. and Wolf M. 1996. Inbreeding avoidance in animals. Trends in Ecology & Evolution *11*(5)**:** 201–6.

Quinn J. S., Woolfenden G. E., Fitzpatrick J. W., and White B. N. 1999. Multi-Locus DNA Fingerprinting Supports Genetic Monogamy in Florida Scrub-Jays. Behavioral Ecology and Sociobiology *45*(1)**:** 1–10.

Qvarnström A. and Bailey R. I. 2009. Speciation through evolution of sex-linked genes. Heredity *102*(1)**:** 4–15.

R Core Team 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reed D. H. and Frankham R. 2003. Correlation between fitness and genetic diversity. Conservation Biology *17*(1)**:** 230–237.

Rice W. R. 1984. Sex chromosomes and the evolution of sexual dimorphism. Evolution *38*(4)**:** 735–742.

Roldan and Gomendio 1999. The Y chromosome as a battle ground for sexual selection. Trends in Ecology & Evolution *14*(2)**:** 58–62.

Rubin B. E. R., Ree R. H., and Moreau C. S. 2012. Inferring phylogenies from RAD sequence data. PLoS One *7*(4)**:** e33394.

Sabeti P. C., Schaffner S. F., Fry B., Lohmueller J., Varilly P., *et al.* 2006. Positive natural selection in the human lineage. Science *312*(5780)**:** 1614–20.

Schluter D., Price T. D., and Rowe L. 1991. Conflicting selection pressures and life history trade-Offs. Proceedings of the Royal Society B *246*(1315)**:** 11–17.

Schueler M. G., Higgins A. W., Rudd M. K., Gustashaw K., and Willard H. F. 2001. Genomic and genetic definition of a functional human centromere. Science *294*(5540)**:** 109–15.

Senn H., Ogden R., Cezard T., Gharbi K., Iqbal Z., *et al.* 2013. Reference-free SNP discovery for the Eurasian beaver from restriction site-associated DNA paired-end data. Molecular Ecology *22*(11)**:** 3141–50.

Shetty S., Griffin D. K., and Graves J. A. 1999. Comparative painting reveals strong chromosome homology over 80 million years of bird evolution. Chromosome Research *7*(4)**:** 289–95.

Shimizu K. and Tsuda K. 2011. SlideSort: all pairs similarity search for short reads. Bioinformatics *27*(4)**:** 464–70.

Sing T., Sander O., Beerenwinkel N., and Lengauer T. 2005. ROCR: visualizing classifier performance in R. Bioinformatics *21*(20)**:** 3940–1.

Skaletsky H., Kuroda-Kawaguchi T., Minx P. J., Cordum H. S., Hillier L., *et al.* 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature *423*(6942)**:** 825–37.

Slate J. 2005. Quantitative trait locus mapping in natural populations: progress, caveats and future directions. Molecular Ecology *14*(2)**:** 363–79.

Smith C. A., Roeszler K. N., Ohnesorg T., Cummins D. M., Farlie P. G., *et al.*

2009b. The avian Z-linked gene *DMRT1* is required for male sex determination in the chicken. Nature *461*(7261)**:** 267–71.

Smith C. A., Roeszler K. N., and Sinclair A. H. 2009a. Genetic evidence against a role for W-linked histidine triad nucleotide binding protein (*HINTW*) in avian sex determination. The International Journal of Developmental Biology *53*(1)**:** 59–67.

Smith C. A. and Sinclair A. H. 2004. Sex determination: insights from the chicken. BioEssays *26*(2)**:** 120–32.

Sobel E., Papp J. C., and Lange K. 2002. Detection and integration of genotyping errors in statistical genetics. American Journal of Human Genetics *70*(2)**:** 496–508.

Steemers F. J. and Gunderson K. L. 2007. Whole genome genotyping technologies on the BeadArray platform. Biotechnology Journal *2*(1)**:** 41–9.

Stiglec R., Ezaz T., and Graves J. A. M. 2007. Reassignment of chicken W chromosome sequences to the Z chromosome by fluorescence in situ hybridization (FISH). Cytogenetic and Genome Research **116:** 132–134.

Stinchcombe J. R. and Hoekstra H. E. 2008. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity *100*(2)**:** 158–70.

Stith B. M., Fitzpatrick J. W., Woolfenden G. E., and Pranty B. 1996. *Metapopulations and wildlife conservation*, Chapter Classification and conservation of metapopulations: A case study of the Florida scrub jay, pp. 187–215. Washington D.C., USA: Island Press.

Stringham H. M. and Boehnke M. 1996. Identifying marker typing incompatibilities in linkage analysis. American Journal of Human Genetics *59*(4)**:** 946–50.

Taylor S. A., White T. A., Hochachka W. M., Ferretti V., Curry R. L., *et al.* 2014. Climate-mediated movement of an avian hybrid zone. Current Biology *24*(6)**:** 671–6.

Thompson E. A. 1994. Monte Carlo programs for pedigree analysis: 1990–1993. Technical report no. 267, Department of Statistics, University of Washington.

Townsend A. K., Bowman R., Fitzpatrick J. W., Dent M., and Lovette I. J. 2011. Genetic monogamy across variable demographic landscapes in cooperatively breeding Florida scrub-jays. Behavioral Ecology *22*(3)**:** 464–470.

Turner W. R., Wilcove D. S., and Swain H. M. 2006. Assessing the effectiveness of reserve acquisition programs in protecting rare and threatened species. Conservation Biology *20*(6)**:** 1657–69.

Väli U., Brandström M., Johansson M., and Ellegren H. 2008. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. BMC Genetics **9:** 8.

van Dongen S. M. 2000. Graph clustering by flow simulation. Ph. D. thesis, University of Utrecht.

Vilà C., Sundqvist A.-K., Flagstad Ø., Seddon J., Björnerfeldt S., *et al.* 2003. Rescue of a severely bottlenecked wolf (*Canis lupus*) population by a single immigrant. Proceedings of the Royal Society B *270*(1510)**:** 91–7.

Wahlberg P., Strömstedt L., Tordoir X., Foglio M., Heath S., *et al.* 2007. A high-resolution linkage map for the Z chromosome in chicken reveals hot spots for recombination. Cytogenetic and Genome Research **117:** 22–29.

Warren W. C., Clayton D. F., Ellegren H., Arnold A. P., Hillier L. W., *et al.* 2010. The genome of a songbird. Nature *464*(7289)**:** 757–62.

White T. A., Perkins S. E., Heckel G., and Searle J. B. 2013. Adaptive evolution during an ongoing range expansion: the invasive bank vole (*Myodes*

glareolus) in Ireland. Molecular Ecology *22*(11)**:** 2971–85.

Wilcoxen T. E., Boughton R. K., and Schoech S. J. 2010. Selection on innate immunity and body condition in Florida scrub-jays throughout an epidemic. Biology Letters *6*(4)**:** 552–4.

Wilcoxen T. E., Bridge E. S., Boughton R. K., Rensel M. A., James Reynolds S., *et al.* 2011. Parental, social and environmental factors associated with hatching failure in Florida Scrub-Jays *Aphelocoma coerulescens*. Ibis *153*(1)**:** 70–77.

Willi Y., Van Buskirk J., and Hoffmann A. A. 2006. Limits to the Adaptive Potential of Small Populations. Annual Review of Ecology, Evolution, and Systematics **37:** 433–458.

Willing E.-M., Hoffmann M., Klein J. D., Weigel D., and Dreyer C. 2011. Paired-end RAD-seq for *de novo* assembly and marker design without available reference. Bioinformatics *27*(16)**:** 2187–93.

Woolfenden G. E. and Fitzpatrick J. W. 1984. *The Florida Scrub Jay - Demography of a cooperative-breeding bird*. Princeton, NJ, USA: Princeton University Press.

Woolfenden G. E. and Fitzpatrick J. W. 1991. *Bird Population Studies*, Chapter Florida Scrub Jay ecology and conservation, pp. 542–565. Oxford University Press.

Woolfenden G. E. and Fitzpatrick J. W. 1996. *The Birds of North America online*, Volume 228, Chapter Florida Scrub-Jay (*Aphelocoma coerulescens*). Cornell Lab of Ornithology.

Young A., Boyle T., and Brown T. 1996. The population genetic consequences of habitat fragmentation for plants. Trends in Ecology & Evolution *11*(10)**:** 413–8.